TELECOMUNICACIÓN

Campus Sur

POLITÉCNICA

# ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA Y SISTEMAS DE TELECOMUNICACIÓN

# PROYECTO FIN DE GRADO

**TÍTULO:** Synthesis of Nature Sounds for Speech Masking

**AUTOR:** Jorge Mir Álvarez

**TITULACIÓN:** Sonido e Imagen

**TUTOR (o Director en su caso):**

Maximilian Schmitt
Aristotelis Hadjakos

**UNIVERSIDAD:** Hochschule für Musik Detmold

**CENTRO:** Erich-Thienhaus-Institut

**PAÍS:** Alemania

**Fecha de lectura:**

**Calificación:**

**El Coordinador de Movilidad,**



Hochschule Für Musik Detmold

Erich-Thienhaus-Institut

Bachelor Thesis

# Synthesis of Nature Sounds for Speech Masking

*Author:*
Jorge Mir Álvarez

*Supervisors:*
Maximilian Schmitt
Aristotelis Hadjakos

July 2014

# *Acknowledgements*

There are a lot of people that have been crucial either directly or indirectly to the completion of this thesis that I would like to thank. I hope you will understand that I address them each in their language.

First and foremost I wanted to thank my two advisors, Maximilian Schmitt and Aristoteles Hadjakos, without whom this thesis would literally not have been completed. Their corrections, suggestions, and patience have been invaluable to my work over the past few months. This has been a challenging learning experience for me, but I know I have come out wiser and more capable. *Vielen Dank, Max und Telis!*

I also need to thank my parents for always giving me incredible opportunities, for always encouraging and supporting me, and for allowing me to have spent this year abroad. Their support in the final stretch of completing this work has been crucial to it actually getting done. Thank you also to my sisters, Tania and Celia, for looking up to (and after) me and providing their encouragement no matter what.

*Gracias a mis abuelos, por haber creído siempre que soy el más listo, el más guapo y el más bueno. Aunque no sea verdad, su confianza en mí siempre ha servido para que diera lo mejor de mí.*

Thank you, Francesca, for always pushing me to be the best version of myself, whether you intended to or not. I'm incredibly lucky and grateful to have you in my life. Thank you, David and Becky, for the brief, sporadic, yet unbelievably entertaining conversations. Thank you, Mirza, for always being there, despite the distance. Thank you to the entire *Unrecorded* staff for giving me a place to distract myself from working on my thesis, and to Frank specifically for having allowed me to lead.

*Quería dar las gracias a Jaime, por las distracciones, todo el apoyo y los ánimos que mandaba desde Bonn. A Julia, también por los ánimos y momentos de diversión. A todos los amigos de la ETSIST, por haber estado ahí para resolver dudas técnicas y no tan técnicas durante estos meses. A Irene por haberme ayudado a dar el último empujón. Y a Raúl por enseñarme que no existen problemas, sólo oportunidades.*

*Letztendlich möchte ich allen meinen Kollegen von HfM und ETI meinen Dank aussprechen, die dieses Jahr zu einer sehr schönen Zeit gemacht haben.*

# Resumen

La introducción de las oficinas abiertas en los años 60 tenía como objeto flexibilizar el ambiente laboral, hacerlo más eficiente y que estuviera más orientado al trabajo en equipo. Como consecuencia, subió el nivel de ruido de fondo, que no sólo dificulta la concentración, sino que causa problemas fisiológicos, como el aumento del estrés, además de reducir la privacidad. Hay estudios que prueban que las conversaciones de fondo en particular tienen un efecto negativo en el nivel de concentración y disminuyen el rendimiento de los trabajadores. Por lo tanto, reducir la inteligibilidad del habla es uno de los principales objetivos en la actualidad.

Un método empleado para hacerlo ha sido el uso de ruido enmascarante, que consiste en reproducir señales continuas de ruido a través de un sistema de altavoces que enmascare el habla. Aunque diversos estudios demuestran que es un método eficaz, los ruidos utilizados hasta la fecha (normalmente ruido rosa filtrado), no son muy bien aceptados por los usuarios. El proyecto colaborativo "*Private Workspace*", dentro del cual se engloba el trabajo realizado en este Proyecto Fin de Grado, tiene por objeto desarrollar un sistema de ruido enmascarador acoplado y adaptativo, además de una estructura física, para su uso en oficinas abiertas con el fin de combatir los problemas descritos anteriormente.

Existen indicios de que los sonidos naturales son mejor aceptados, en parte porque pueden tener una estructura física que simule ser la fuente de los mismos. La utilización de grabaciones directas de estos sonidos no está recomendada por varios motivos, y por lo tanto los sonidos naturales deben ser sintetizados. El presente trabajo consiste en la síntesis de una textura de sonido (en inglés *sound texture*) para ser usada como ruido enmascarador, además de su evaluación. La textura está compuesta de dos partes: un sonido de viento sintetizado mediante síntesis sustractiva y un sonido de hojas sintetizado mediante síntesis granular. Diferentes combinaciones de estos dos sonidos producen cinco variaciones de ruido enmascarador. Estos cinco ruidos han sido evaluados a diferentes niveles, junto con ruido blanco y ruido rosa, mediante una versión modificada de un *Oldenburger Satztest* para comprobar cómo afectan a la inteligibilidad del habla, y mediante un cuestionario para una evaluación subjetiva de su aceptación. El objetivo era encontrar qué ruido de los que se han sintetizado funciona mejor como enmascarador del habla.

El proyecto consiste en una introducción teórica que establece las bases de la percepción del sonido, el enmascaramiento psicoacústico, y la síntesis de texturas de sonido. Se explica a continuación el diseño de cada uno de los ruidos, así como su implementación en MATLAB. Posteriormente se detallan los procedimientos empleados para evaluarlos. Los resultados obtenidos se analizan y se extraen conclusiones. Por último, se propone un posible trabajo futuro y mejoras al ruido sintetizado.

# Abstract

The introduction of open-plan offices in the 1960s with the intent of making the workplace more flexible, efficient, and team-oriented resulted in a higher noise floor level, which not only made concentrated work more difficult, but also caused physiological problems, such as increased stress, in addition to a loss of speech privacy. Irrelevant background human speech, in particular, has proven to be a major factor in disrupting concentration and lowering performance. Therefore, reducing the intelligibility of speech and has been a goal of increasing importance in recent years.

One method employed to do so is the use of masking noises, which consists in emitting a continuous noise signal over a loudspeaker system that conceals the perturbing speech. Studies have shown that while effective, the maskers employed to date – normally filtered pink noise – are generally poorly accepted by users. The collaborative "Private Workspace" project, within the scope of which this thesis was carried out, attempts to develop a coupled, adaptive noise masking system along with a physical structure to be used for open-plan offices so as to combat these issues.

There is evidence to suggest that nature sounds might be more accepted as masker, in part because they can have a visual object that acts as the source for the sound. Direct audio recordings are not recommended for various reasons, and thus the nature sounds must be synthesized. This work done consists of the synthesis of a sound texture to be used as a masker as well as its evaluation. The sound texture is composed of two parts: a wind-like noise synthesized with subtractive synthesis, and a leaf-like noise synthesized through granular synthesis. Different combinations of these two noises produced five variations of the masker, which were evaluated at different levels along with white noise and pink noise using a modified version of an Oldenburger Satztest to test for an affect on speech intelligibility and a questionnaire to asses its subjective acceptance. The goal was to find which of the synthesized noises works best as a speech masker.

This thesis first uses a theoretical introduction to establish the basics of sound perception, psychoacoustic masking, and sound texture synthesis. The design of each of the noises, as well as their respective implementations in MATLAB, is explained, followed by the procedures used to evaluate the maskers. The results obtained in the evaluation are analyzed. Lastly, conclusions are drawn and future work is and modifications to the masker are proposed.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Open-plan offices were introduced in the 1960s with the idea of being flexible and efficient, as well as encouraging team-oriented work [52]. However, despite these supposed advantages, this concept also had a major downside: the amount of people in the room raised the noise floor level, making concentrated work more difficult. This has resulted in a diversion of attention, a decrease in concentration, an increase in mistakes made, a disturbance in speech intelligibility, as well as a general increase of stress hormones, with the resulting psychosomatic consequences [9].

These disadvantages have negated the initially sought benefits of open-plan offices, with recent trends opting for combinations between the styles. However, none of these have solved the acoustic problems common in many offices.

Experiments have shown that human speech in particular causes more of a distraction than the noise that, for instance, comes from any of the technical machines used in the room. Bradley and Gover [6] found that speech intelligibility correlates with the subjective ratings on the perceived disturbance. That is, that the higher the intelligibility of the ambient or background speech, the higher the subjective disturbance ratings are. Some published studies even seem to indicate that the perception of irrelevant speech - even if the content is not understood - is enough to produce a significant decrease in concentration and memory [3]. This proved to also be true for foreign languages and speech-like noises [22].

Because of these adverse effects, reducing background speech level and/or its intelligibility has been a common goal in recent years. One solution has been the addition of physical constructs such as screens, windows, or walls to separate offices. Additional recommended measures recommended by some include adding absorptive material to different surfaces [30]. Still, these barriers have often not been enough to reduce speech audibility or even intelligibility [36].

To this effect, masking systems were employed with the goal of obtaining low or even

negative values of speech-to-noise ratio[1] [15]. These systems produce a continuous noise signal over a loudspeaker system so as to conceal the perturbing speech so that it can no longer be perceived and is therefore no longer a distraction. The downside to this solution is that it can further raise the noise floor level, leading to what is known as the Lombard effect or reflex, which describes the tendency of the speaker to speak louder as the noise level around them increases, in turn raising the noise floor level, producing a cycle [23].

Given this, the goal of the Private Workspace project is to develop a coupled, adaptive noise masking system, in addition to a physical construct, for open-plan offices that counters these effects. Each of the four partners of the project - *Hochschule für Musik Detmold* (HfM), *Hochschule Ostwestalen-Lippe* (HS-OWL), SilenceSolutions GmbH (SilenceSolutions), and SINUS Messtechnik GmbH (SINUS Messtechnik) - have a clearly defined series of tasks that joined together will produce a system that solves the aforementioned problems while being accepted by users. The full details of how this will be achieved are beyond the scope of this thesis and held to a confidentiality agreement, but the specific objectives of this thesis will be discussed in section 1.3.

## 1.1 The Irrelevant Speech Effect (ISE)

The Irrelevant Speech Effect (ISE) describes the phenomenon that immediate verbal short-term memory performance (also known as serial-recall) is significantly reduced when exposed to irrelevant speech and/or certain non-speech sounds, even if participants have been told to ignore them [16]. In addition, the level of the speech seemed to not have any influence on the ISE within the normal comfortable range of hearing (below 80 dBA), and a level variation between the speech was found to not cause any additional disruption either [31].

While the exact nature of the ISE is still not entirely well known, the characteristics of the background sound that provoke it are. The essential feature is the presence of distinct temporal-spectral variations that allow for the perceptual segmentation of the irrelevant sound, while successive perceptual tokens vary in acoustic-perceptive perspective [36]. In other words, a sound that has changing-state characteristics has proven to be more disruptive to short-term memory performance.

This not only explains why speech is so disruptive to concentration – it can be perceptually segmented and is a signal in which successive perceptual tokens vary significantly – but also describes the necessary conditions for a speech masker – one that does not possess changing-state characteristics.

[1] The speech-to-noise ratio $L_{SN}$ is described as the level-difference between speech and background noise

Alternatively, the ISE also describes what can be considered a primary goal for maskers: rather than completely conceal the background speech, it is necessary to reduce the number of spectrally distinctive features and the spectrally distinctive features of the speech, i.e. its intelligibility [31].

## 1.2 Related Works

The need for masking noises in office settings was recognized in even the early stages of the open-office concept [17, 49], though its use has not been extended, perhaps due to the little scientific research done in the field and the contradictory results [18]. In addition, there is little specific information about the characteristics of the noise that would prove to be most satisfactory, both in objective performance and subjective acceptance. Up to the year 2002, there had been no systematic research published on which to base the spectral content or level of the noise to be used as a masker [45]. Since then some advances has made, but without any definite conclusions.

Most studies that have been carried out have used filtered noise as a speech masker [45, 36, 46]. Under laboratory conditions, Veitch et al. [45] found that the optimum spectrum for a noise masker was similar to the sound spectrum, in which sound pressure level is reduced 5 dB per octave in the frequency range of 100 to 10000 Hz, obtained by filtering pink noise. Music has also been used by some workers to mask noise, however, the higher the concentration the task requires, the more disturbing background music is [28].

Haapakangas et al. [15] state as recently as 2011 that current literature does not directly conclude what kind of masking sound would be most effective for open-plan offices in terms of work performance and acoustic satisfaction. Though they suggest that nature sounds have been a "particular interest", to date there has been little evidence of scientific research work done using nature sounds as maskers.

Because they can be considered as naturally occurring and therefore have a clearly (visually) identifiable physical source, these types of sounds might be perceived as less disturbing than the previously employed filtered noise, which is not present in any form in nature. Direct sound recordings, however, might not be ideal as maskers due to the possible unpredictability of sonic events and patterns (i.e. birds chirping or other animal noises, irregular or nonexistent wind patterns, etc). Changing-state characteristics such as these have proven to be disruptive to memory performance, explained in section 1.1. Another problem might be the possible perception of repetitive patterns as the audio recordings are looped (obtaining an extended recording without any unwanted artifacts could prove to be very challenging), which might be considered distracting. It seems logical to suggest that these problems might be best solved by using synthesis to create a natural noise that could then be used as

a masker. A synthesized signal will only contain the desired sonic elements, and, if implemented correctly, should allow to have an arbitrarily long signal without repetition.

Some studies have already been carried out using natural maskers. Haapakangas et al. [15] found that, under their testing conditions (a serial recall task, a creative thinking task, and a proofreading task), spring water sound generated "significant improvement in objectively measured performance compared to speech" and "had the most benefits in terms of both subjective and objective indicators". Schmitt [37] used (among others) a forest and a waterfall sound as maskers, though he did not arrive at any significant conclusions.

It is also interesting to note that in a study comparing the effects of verbal versus visual information about sound sources on the perception of environmental sounds, Abe et al. [1] conclude that "the influence of visual and verbal information on the auditory evaluation of environmental sounds is considerable". Therefore, another possible advantage to using a natural sound masker could be the possibility of having a visual "source" of the sound — even if this visual source is not the actual source of the masker.

## 1.3 Objectives

As has been outlined in the previous sections, there is still significant research to be done in the realm of using nature sounds as speech maskers. The findings by Abe et al. [1] mentioned in the previous section, suggest that a natural sound noise might be more successful if paired with a visual source, which will be the limiting factor

In this case, a physical system has been developed by the HS-OWL to simulate the fluttering of leaves in the wind. An exciter emitting a low frequency outside the range of human hearing (10 Hz) to lightly move a plastic panel that has leaf-like shapes hanging from it. The idea is to pair this structure with the noise these "leaves" would make in the wind, and which would serve as the speech masker.

The natural masker, which will be synthesized, must provide a solution to the issues described in sections 1.1 and 1.2, as outlined below:

- It must mask human speech sufficiently well so as to decrease the perceived speech information and therefore increase speech privacy
- It must be able to be perceived as real, and contain elements of both wind noise and leaves noise, so as to match the aforementioned physical structure
- It must be generally accepted by the end users, who must not find it to be too unpleasant or distracting

- It must be able to be of a previously-defined arbitrary duration without significant changes in its spectral or temporal characteristics so as to avoid the Irrelevant Speech Effect and therefore cause minimal disruption to concentration

The main objective of this Bachelor Thesis is, then, to develop a synthesized natural noise masker that meets the previously outlined conditions, as well as design, carry out, and analyze a listening test that evaluates its effectiveness and user acceptance. These tasks correspond to a portion of those assigned to the HfM within the framework of the Private Workspace project.

## 1.4 Structure of the Thesis

This thesis is structured into five different chapters. This first chapter has served to establish the technological framework of the thesis, including related works and its objectives. Chapter 2 encompasses all of the theoretical background necessary to understand the main work done, including the functioning of the human ear, how we perceive sound, psychoacoustic masking, the characteristics of speech signals, and an overview of sound texture synthesis. Chapter 3 describes how the speech masker was synthesized, including its implementation in MATLAB, and chapter 4 the tests that were designed and carried out to evaluate it, as well as the obtained results. Lastly, chapter 5 draws some conclusions and proposes potential future work.

# Chapter 2

# Theoretical Background

Psychoacoustics is the branch of psychophysics concerned with the perception of sound and the sensations produced by sounds. It *"seeks to reconcile the acoustical stimuli and all the scientific, objective, and physical properties that surround them, with the physiological and psychological responses evoked by them"* [32]. Psychoacoustics focuses on the relationship between the acoustic sound signal and the auditory events associated with it [5].

One of the objectives of psychoacoustics is to quantify these relationships so as to take advantage of the defects and strengths of our hearing. Sound masking seeks to take advantage of these defects and strengths.

## 2.1 The Human Ear

The ear is commonly split into three different regions: the outer ear, the middle ear, and the inner ear. Figure 2.1 shows a schematic drawing of each of these parts. The outer ear's purpose is to collect the sound energy present in a field and channel it through the outer ear canal and into the ear drum. The outer ear canal has a strong influence on the frequency response of our hearing organ; the system can be compared to that of an open pipe with a length of about 2 cm, nearly equivalent to that of a quarter of the wavelength of frequencies around 4 kHz. This results in an increased sensitivity of our hearing around this range [11].

The middle ear's primary function is to transform the oscillation of air particles that occurs in the outer ear into the oscillation of the fluids that excite the sensory cells in the inner ear. To avoid large losses of energy through reflections, a transformation occurs in the middle ear to match the impedances of the air outside and liquid inside [11]. To this end, the middle ear acts like a lever. The ear drum serves as a pressure receiver and operates over a broad frequency range. The motion produced in the ear

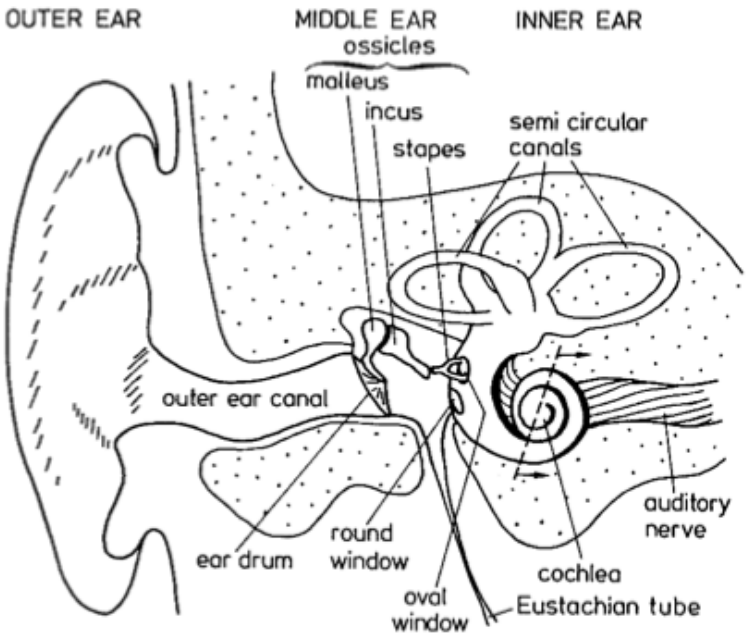



Figure 2.1: Schematic drawing of the external, middle, and inner ear, from [11]

drum are transmitted to the stapes by the ossicles, which consists of the malleus, incus, and stapes, all made of very hard bone. The stapes footplate and the ring-shaped oval window are the entrance to the inner ear. Along with the a lever ratio[1] of about 2, the middle ear also produces a transformation depending on the ratio of the area of the relatively large ear drum to that of the footplate. This ratio is about 15. Through these two ratios, an almost perfect impedance match is achieve around 1 kHz [11]. Below this frequency the involuntary tightening of the *tensor tympani* and *stapedius* damp the ability of the ear drum and stapes to transmit sound by about 12 dB, with the main objective of reducing the audibility of one's own speech [50].

The inner ear primarily consists of the snail-shaped cochlea, which is embedded into the extremely hard temporal bone, and the auditory nerve. The cochlea is filled with liquid, and consists of three channels or scalae (scala vestibuli, scala media, and scala tympani). The footplate of the stapes is in direct contact with the scala vestibuli. Oscillations are transmitted to the basilar membrane, which resonates through the fluids. The organ of Corti, which contains the important sensory cells or hair cells and is located on the basilar membrane, transforms the mechanical oscillations in the inner ear into a signal that can be processed by the nervous system [11].

The sound we perceive is a result of where exactly the basilar membrane resonates. Low frequencies produce oscillations of the membrane near the helicotrema and high frequencies near the oval window. This oscillation is produced as a traveling wave,

[1]The lever ratio is the ratio of the output force to the input force

beginning with a small vertical displacement near the oval window, reaching a maximum at a certain location, and then dying out in the direction of the helicotrema. A tone will produce a resonance at some specific location along the basilar membrane's 32 mm. If, for example, three tones of sufficiently different frequencies are present at once, they will be separated in the inner ear [11]. This means that the basilar membrane is largely responsible for frequency separation. (Just how sufficiently different the frequencies must be is addressed in section 2.3.1.)

## 2.2 Sound Perception

The sound waves that reach our ears are a result of pressure variations in the air (or another medium we may find ourselves in). These variations are referred to as sound pressure and are measured in Pascal [Pa]. A sound pressure level (SPL), is calculated as shown in (2.1), where $P$ is the given pressure variation and $P_{ref}$ is a reference pressure. A typically used value of this reference pressure is $20\mu Pa$, which corresponds to the minimum pressure variation that a human ear can hear (at 1 kHz).

$$\mathrm{SPL} = 20\ \log_{10}\frac{P}{P_{\mathrm{ref}}} \quad [\mathrm{dB}] \tag{2.1}$$

Human hearing is limited in level by the hearing threshold and the threshold of pain or feeling and in frequency by the physical limits of our hearing organ, in a range typically considered to be between 20 Hz and 20 kHz, though this is age-dependent, with adults closer to an upper limit of 16 kHz. The hearing area, as seen in Figure 2.2 is considered to be the area between these two thresholds. Figure 2.2 also shows the typical areas for both music and speech. For signals that aren't pure tones, *sones* are used as a measure of loudness. The level of a 40 dB, 1 kHz tone was proposed as a reference, i.e. the loudness of this signal is equal to 1 sone [11].

The curves that delimit the threshold of pain and threshold of hearing (or threshold in quiet) are the two limits of what are known as the equal loudness curves or contours. The curves, seen in figure 2.3, indicate the sound pressure level for which a listener perceives a constant loudness when presented with a pure tone. They are measured in *phons*, a unit of loudness level for pure tones. By definition, the number of phon of a sound is the dB SPL of a sound at a frequency of 1 kHz that sounds just as loud. That is, that a sound that is 70 phons means it is as loud as a 70 dB, 1 kHz tone.

The threshold in quiet indicates as a function of frequency the sound pressure level of a pure tone that is just audible [11]. As can be observed in figures 2.2 and 2.3, it is not the same across all frequencies. Already mentioned is the slight dip in the threshold in quiet in the 2 to 5 kHz range due to the physical construct of the

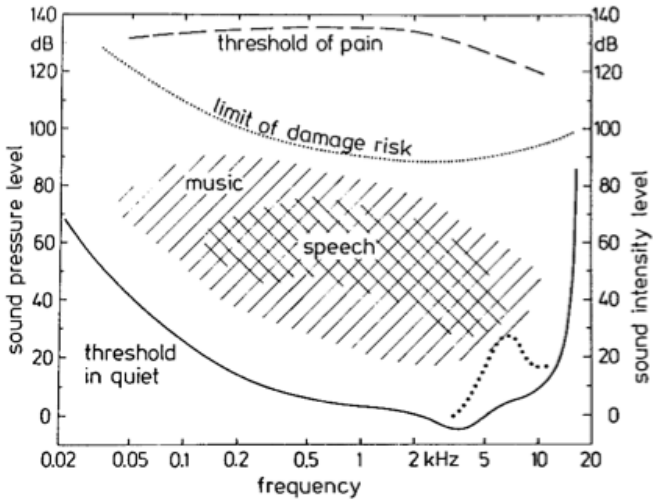



Figure 2.2: Hearing area, i.e. the area between the threshold in quiet and threshold of pain. From [11]

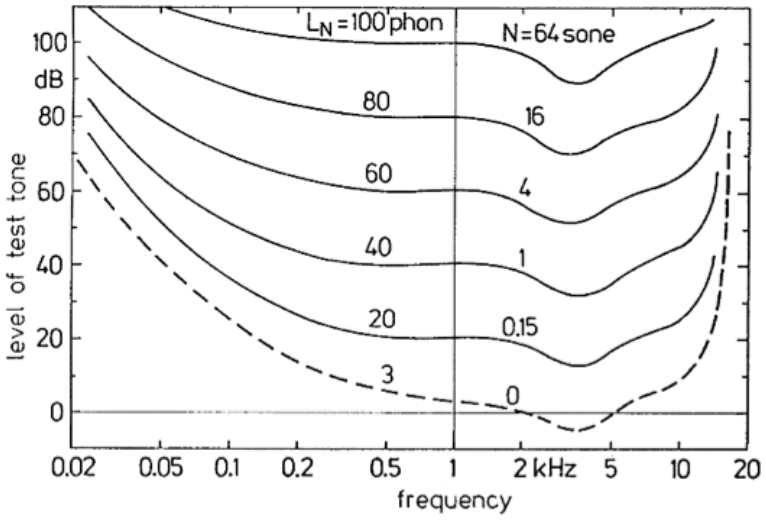



Figure 2.3: Equal-loudness curves for pure tones. $L_N$ is the loudness level and $N$ is the loudness. From [11]

human ear canal. More significant is the sharp increase in higher frequencies; this is the area that limits the limit of hearing, and that is most affected by age. Under normal conditions, between ages 20 and 25 the upper limit is between 16 and 18 kHz, though towards the higher frequencies the threshold in quiet is at very high levels. At age 60 not only is the upper of limit significantly lower (between 12 and 15 kHz), but the threshold in quiet is considerably higher (around 30 dB as opposed to around 10 dB at age 25). At age 40 the threshold shift is about half as much as that for 60 years [11].

This upper range of hearing will be significant when synthesizing the masker, as it will not need to extend to frequencies beyond 16 kHz since these will not be heard at normal levels by the majority of working adults.

## 2.3 Masking

At its most basic, masking occurs when one sound signal is enough to partially or completely cover up, or mask, another. For example, if two people are maintaining a conversation on the street and a loud truck drives by, the sound of the truck might be enough that the conversation is no longer audible. If this is the case there are two possible options: either raising the voice level so as to speak more loudly than the sound of the truck, or waiting until the track has passed to resume the conversation. This kind of masking is called simultaneous masking, and in this particular case is primarily due to level.

If the truck was loud enough that the two individuals could no longer hear each other at all, the masking is known as total masking. More likely, however, would have been partial masking, in which the loudness of the test tone is reduced without being completely masked.

### 2.3.1 Critical Bands

The physical structure of the cochlea and the basilar membrane means that our hearing processes sounds in relatively narrow frequency bands, know as critical bands. These critical bands can be modeled as a series of band-pass filters. At low frequencies, these "filters" have a constant width of 100 Hz, while at frequencies above 500 Hz critical bands have a bandwidth proportional to the frequency, approximately 20% of the center frequency [11]. Each of these bandwidths is known as a critical bandwidth.

Effectively, a critical band is the band of audio frequencies within which a second tone interferes with the perception of a first tone. If two tones within the same critical band are perceived, they are perceived as one sound instead of two. This

phenomenon is what gives way to auditory masking, explained more in depth in the following sections.

### 2.3.2 Masking Threshold

In order to measure the effect of masking in a quantitive way, it is necessary to determine the masking threshold. The masking threshold is the sound pressure level of a test sound (normally a sinusoidal tone), necessary to just be audible in the presence of a masker [11]. The masking threshold almost always lies above the threshold in quiet, and it is equal to the threshold in quiet when the frequencies of the masker and the test sound are very different [11]. Only the masker frequencies corresponding a given critical bandwidth contribute to the masking of the signal.

### 2.3.3 Temporal Effects

The most common phenomenon when referring to masking in time is simultaneous masking. Simultaneous masking occurs when a sound is made inaudible by another of a duration at least equal to it and playing simultaneously [26].

Masking does not necessarily need to be simultaneous. In some cases, "premasking" or "postmasking" can also occur. For the former, the test sound needs to be a short burst or impulse that are presented before the masker is switched on, and typically it is not too effective. The opposite, that is, the test sound being present after the masker is switched off, is more pronounced, and is what is known as "postmasking".

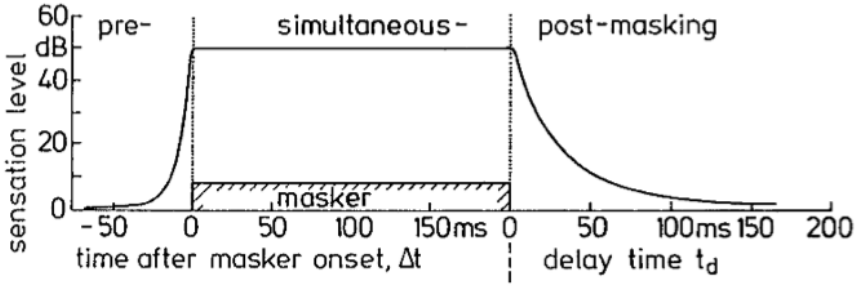



Figure 2.4: Schematic drawing illustrating the regions in which premasking, simultaneous masking, and postmasking occur. From [11]

### 2.3.4 Masking Pure Tones with Noise

Most scientific studies examining masking phenomena have been carried out using pure tones as the test signal, and using either noise or pure tone maskers. For the scope of this thesis, only masking with noise will be discussed.

#### Pure Tones Masked by Broad-Band Noise

Figure 2.5 shows the threshold level of a test tone masked by white noise. As can be seen, the masking is only horizontal at low frequencies, about 17 dB above the given density level. Above 500 Hz, there is an increase of approximately 10 dB per decade. It is interesting to note that there doesn't seem to be a linear relationship between the changes in the threshold of hearing and the masking level.

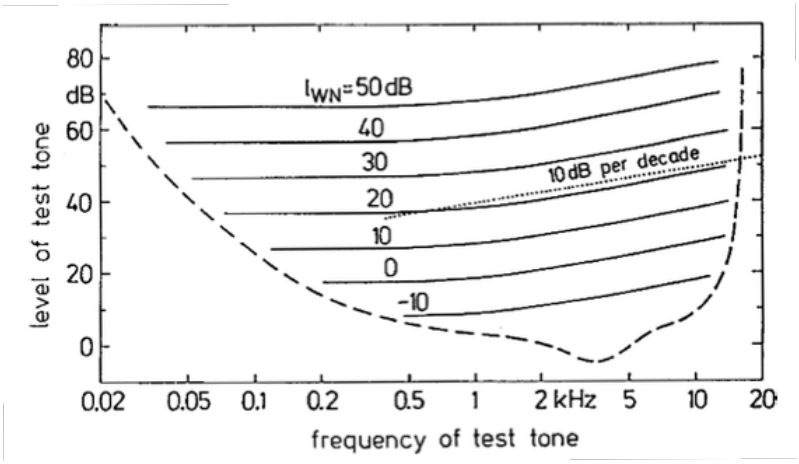



Figure 2.5: Level of test tone just masked by white noise of given density $l_{WN}$, as a function of the test-tone frequency. The dotted line represents the threshold of hearing. From [11].

#### Pure Tones Masked by Narrow-Band Noise

For his explanation, Fastl [11] defines narrow-band noise as "a noise with a bandwidth equal to or smaller than the critical bandwidth". Figure 2.6 shows the thresholds of pure tones masked by critical-band noise at 60 dB and center frequencies of 0.25, 1, and 4 kHz. Figure 2.7 shows the thresholds of pure tones masked by critical band noise with a center frequency of 1 kHz at different levels. From these two figures it it is possible to observe four important phenomena:

1. Masking is not symmetrical around the center frequency. Frequencies higher than the center frequency are more easily masked than those lower.
2. The maximum of the masking thresholds tends to be lower for higher centre frequency-maskers.

3. Masking acts differently in low and high frequencies. As seen in 2.6, the masking that occurs with the critical-band noise centered at 0.25 kHz is much broader than the one that occurs with the noises centered at 1 or 4 kHz.

4. The frequency dependence of the masked threshold is level-dependent. That is, at low levels masking affects a much narrower band than at higher levels. An extreme example is seen in 2.7, where with the 100 dB masker the effect was prominent past 10 kHz.

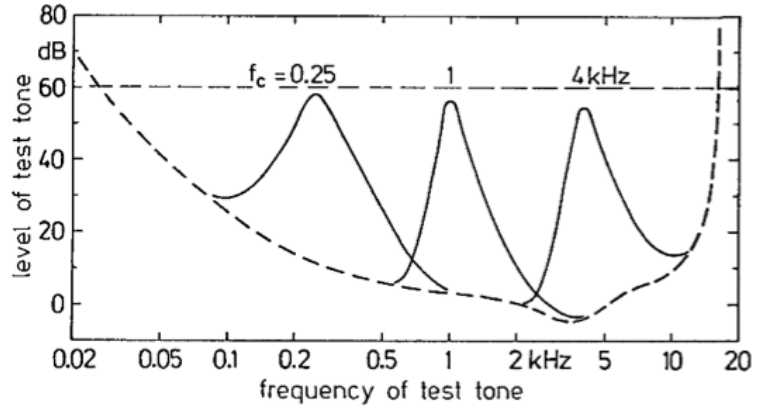



Figure 2.6: Level of test tone just masked by critical-band noise with a level of 60 dB, and center frequencies of 0.25, 1, and 4 kHz. The dotted line represents the threshold of hearing. From [11].

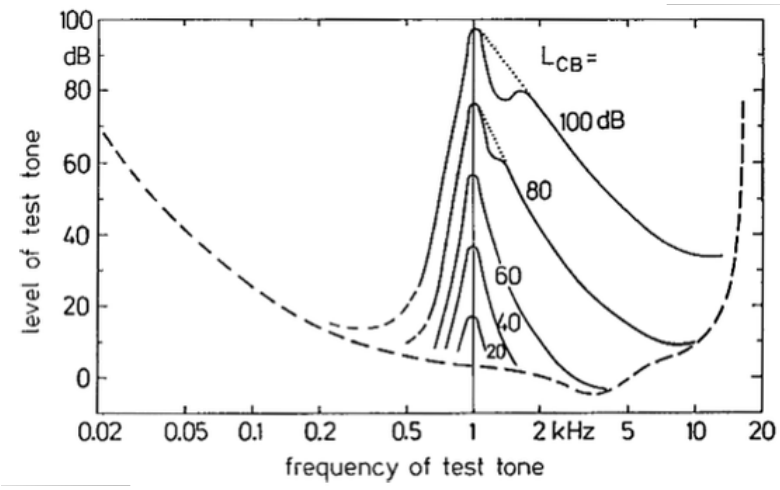



Figure 2.7: Level of test tone just masked by critical-band noise with center frequency of 1 kHz and different levels. The dotted line represents the threshold of hearing. From [11].

### 2.3.5 Partial Masking

A masking sound does not only a produce a shift in the threshold in quiet to the masking threshold, but also produces a masked loudness function that has to be

steeper than the unmasked loudness curve, called a partial masked loudness curve [11]. This curve illustrates the phenomenon that, when a masker is present a partially-masked sound is perceived to be less loud. Partial masking produces a loudness function comparable to what is described by audiologists as *loudness recruitment* [11], in essence a reduction of dynamic range that occurs when the outer hair cells on the organ of Corti are damaged [2].

Partial masking is actually most relevant to the scope of this thesis, since as described in section 1.1 it is not necessary to completely mask the speech signal. Instead, the speech need only be partially masked enough so that it is no longer intelligible. Practically, this means that the level of the masker does not need to be as high as it would need to under total masking conditions.

## 2.4 Characteristics of Speech Signals

In order to know the spectral characteristics of a masker, it is important to be aware of the characteristics of the signal it is expected to mask. In this case, this involves briefly analyzing the human voice and the characteristics of speech signals.

Figure 2.8 shows a cross-sectional view of a human head that illustrates acoustically important features in vocal production. Most of the elements show in the figure form what is known as the vocal tract, which is the air cavity that is used in production of human speech [27].

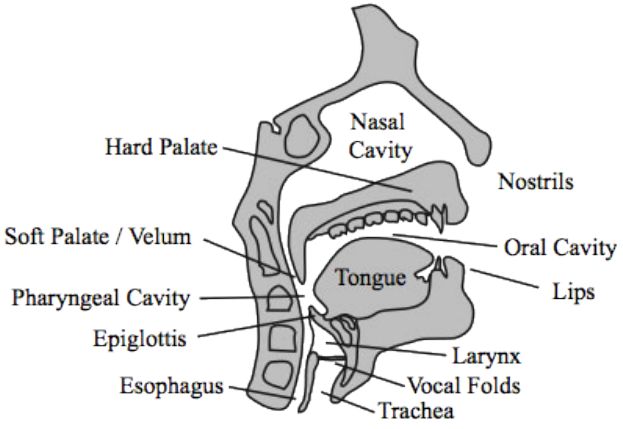



Figure 2.8: The human vocal system, from [27]

A source-filter combination can be used to model the sounds created during continuous speech [10]. The vocal tract serves as a variable shape acoustic filter that changes its resonant properties during speech. The excitation comes as a result of the pressure pulses that are generated as air flowing up from the lungs causes the vocal folds to periodically open and close. The vocal tract filters the glottal source

signal, imposing its spectral characteristics, and the resulting output from the lips is perceived as speech sounds [27].

The space in between the vocal folds is called the glottis. During speech-generation, a contraction of the diaphragm muscles forces a steady stream of air up from the lungs, which builds up behind the closed glottis. When the force against the vocal folds is greater than the elastic tension holding them together, they move apart, briefly release a pulse of air. A posterior reduction in pressure and the tension in the vocal folds will then cause the glottis to abruptly return to its closed state [27]. This periodic cycle is what results in the fundamental frequency of the speech signal, which is ranges from 85 to 180 Hz for adult males, and from 165 to 255 Hz for adult females [43]. These fundamental frequencies will serve as reference points as to how far the masker needs to extend in the low end of the frequency spectrum. A direct corollary, then, is that a subwoofer is *not* required to emit the masker.

In simplified terms, the vocal tract acts a tube for the air pulses generated in the glottis, which produces resonances for certain frequencies, called formants [27]. Each vowel sound, and vocal tract, produces different formants, which in turn gives vowels and speakers their characteristic sound. Figure 2.9 shows averages formant frequencies and their corresponding bandwidths (in parentheses) from a range of vowels occurring in natural speech.

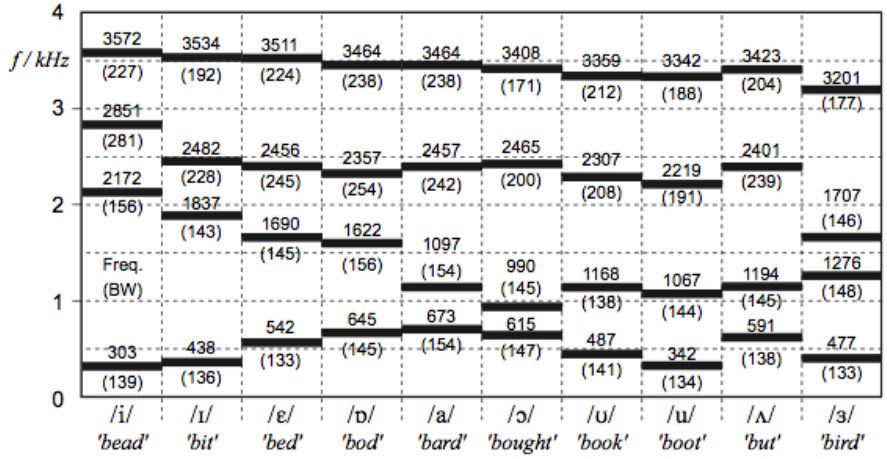



Figure 2.9: Average formant frequencies and bandwidths for male speakers, from [27]

While figure 2.9 might seem to give an idea of how far the masker needs to extend in frequency in order to reduce the intelligibility of speech, it only illustrates vowel sounds. Consonants are created by short and abrupt articulator movements[2], which gives rise to noise-like excitation, called frication [27]. This impulsive, noise-like excitation results in a broad spectrum, which can extend to around 10 kHz [13] and must also be masked. If the fundamental frequency sets the lower limit for

[2]The articulators are the tongue, jaw, teeth, and lips.

where the masker should reach in frequency, the impulsive consonants set the upper limit.

It is also important to note a few other key characteristics of speech signals. Vowels generally represent the loudest sounds, compared to the relatively faint consonants [11], therefore, more energy will be needed in the range that affects vowel sounds. In addition, it is also known that middle and higher frequency portions of speech signals are decisive cues for segmentation, and that therefore consonants are essential for speech intelligibility. Specifically, frequencies between 1000 and 4000 Hz are especially important for understand speech, and are therefore the parts of speech that contribute to the Irrelevant Speech Effect [36].

## 2.5 Sound Texture Synthesis

### 2.5.1 Definition of a Sound Texture

Though no single definition exists, a sound texture is generally understood to be a sound that is composed of several micro-events, but whose features are stable over a longer period of time [39]. A few examples of real sound textures could be the sound of fire, a waterfall, or traffic noise.

Perhaps the best way to understand a sound texture is through the analogy given by Saint-Arnaut and Popat [35]: *"A sound texture is like wallpaper: it can have local structure and randomness, but the characteristics of the fine structure must remain constant on the large scale."* In the particular case of the objectives of this thesis it becomes particularly relevant, as the goal is that the generated noise is as easy to ignore as wallpaper.

Figure 2.10 shows the relationship of the potential information content of music or speech, sound textures, and noise over time. For as long as they go on, speech or music continue to provide new information, and noise generally contains very little information. Sound textures, however, by definition must add no relevant information with time so as to maintain their wallpaper-like quality but must contain a minimum of relevant information so that they are useful. This characteristic is especially important when considering the Irrelevant Speech Effect discussed in section 1.1, as it provides the necessary absence of changing-state features that would make the synthesized masker less of an impairment to performance.

Saint-Arnaut and Popat arrive at the following definition of a sound texture:

1. Sound textures are composed of basic sound elements, called atoms;
2. atoms occur according a higher-level patter, which can be periodic, random, or a combination of the two;

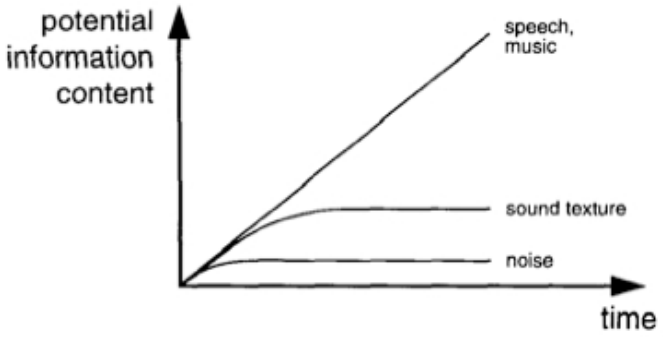



Figure 2.10: Potential information content vs. time, from [35]

3. there can be no complex message, that is, the high-level characteristics must remain constant over long periods of time;
4. the high-level pattern must be completely exposed with the "attention span" of a few seconds. Saint-Arnaut and Popat refer to the "attention span" of a sound texture as "the maximum time between events before they become distinct";
5. high-level randomness is acceptable provided there are enough occurrences within the attention span to make a good example of the random properties.

As Strobl et al. indicate [42], the main objective of the resynthesis of a sound texture is to create a new sample longer in duration that maintains similar qualities to the original. This ability to be stretched and expanded is another quality a sound texture should have, and directly ties in with the necessary condition that no new information is exposed over time. The expansion, without repeatability, is particularly relevant for the scope of this thesis.

A *soundscape* is the sum of a several sounds to compose an entire scene, where some of the sounds could be sound textures [39]. For example, a soundscape could be a "thunderstorm", formed by several different sounds and sound textures, such as thunder, rain, strong wind, and maybe trees/leaves rustling in the wind. In this example, the rain, strong wind, and rustling trees/leaves sounds would most likely be considered sound textures while the thunder would not, as the information content does not remain constant over time. For the objectives of this thesis, what will be synthesized is a relatively simple soundscape consisting of two different sound textures: a leaf noise and a wind noise.

### 2.5.2 Survey of Synthesis Methods

Schwarz [39] proposes that there exist two different main uses of sound texture synthesis:

**Expressive texture synthesis:** The aim is to interactively generate sound for musical purposes, be they composition, performance, or sound art. In this case a sound texture serves to differentiate the generated sound material from tonal and percussive sound. In other words, a sound texture is predominantly defined by timbre rather than pitch of rhythm.

**Natural texture resynthesis:** The goal is to synthesize environmental or human textural sound, possibly to be used as part of a larger soundscape. A certain degree of realism is sought, but in most cases *credible texture synthesis* can be sufficient, in that the textures convey the desired ambience or information.

Judging by these two potential objectives, it is clear that the second is more relevant, and methods that refer to this goal will be the focus of the survey. To further classify the sound texture synthesis methods, the classification proposed by Misra and Cook [25], in which they distinguish between "synthesis from scratch" and "synthesis from existing sounds" will be used.

### Synthesis From Scratch

Synthesis from scratch refers to the replication of real-world sounds using physical or perceptual models, without the raw material of existing audio samples [25]. In this case, the model used might represent the sound's source or environment or the perceptual characteristics desired. The main advantage of using these methods is the possibility of having a high level of control over the resulting sound.

One major type of method that can be classified as synthesis from scratch is *physical modeling*. Physical modeling uses a previously defined mathematical model to generate the waveform of the sound to be synthesized. A set of equations and/or algorithms is used to simulate physical sound sin the real world [34]. One type of physical modeling uses modal resonance models to recalculate inexpensively synthesis able modes from expensive rigid body simulations [44]. There are other methods which are physically informed, meaning that they control the signal models by the output of a physical model that captures the behavior of the sound source [7].

As no real mathematical or physical models of leaf and/or wind sounds currently exist, or at least are not widely/publicly available, synthesis from scratch methods, or more specifically using physical modeling does not seem to be the most appropriate mean by which to synthesize the desired masker.

### Synthesis From Existing Sounds

As its name indicates, methods classified as synthesis from existing sounds use audio material as a source for the synthesis. The source can either be used directly for

the synthesis itself by rearranging samples in the time-domain (concatenative techniques) or analyzed for its spectral characteristics and then synthesized. Of course, a combination of the two is also possible.

**Concatenative techniques**

Schwarz [38] describes concatenative synthesis methods as those that use a "large database of source sounds, segmented into *units*, and a *unit selection* algorithm that finds the sequence of units that match best the sound or phrase to be synthesized, called the *target*". These units are then modified as needed, and then concatenated in the time domain, potentially using a cross-fade. Specifically he focuses on *corpus-based concatenative synthesis*, which makes it "possible to create a sound by selecting snippets from the large database (the corpus) by navigating through a space where each snippet is placed according to its sonic characters in terms of audio descriptors" [39]. While this approach seems interesting, two main problems exist: 1) due to a limited amount of high-quality material it might be difficult to build the necessary large database of samples, and 2) since corpus-based concatenative synthesis is such a novel development, there still isn't much research concluding how effective it is.

Instead, it might be wise to look at the precursor to corpus-based synthesis. *Granular synthesis* involves generating thousands of relatively short *sonic grains* to form larger acoustic events [33]. It was initially proposed by physicist Dennis Gabor in conjunction with a theory of hearing. Gabor referred to *acoustical quanta*, whose representation could be used to describe any sound [12], and which was later proven by Bastiaans [4]. Gabor suggested organizing the grains into events, characterized by 12 parameters, such as duration, initial waveform, initial center frequency, bandwidth, initial grain density and others. In his book *Formalized Music*, Iannis Xenakis proposed a compositional theory with the sound grains, describing a possible approximation to Gabor's model in the context of analog synthesis [51].

The sound grains used in the synthesis can either be created from scratch, or obtained by splitting an audio sample into small segments. The goal is to obtain a variation on a signal that still bears a significant resemblance to the original [20]. Hoskinson and Pai [20] argue that "a long audio sample is not even required; it suffices to specify the shape of the grain and its envelope". In the type of granular synthesis known as *granulation*, in which an audio sample is used [33], a grain is only an arbitrary slice chosen independently of the sound's inherent structure [20]. The length of the grain used determines how much of the temporal envelope of the original source is maintained. When using short grains, the result is a very pulse-like signal, whereas longer grains allow to maintain some of the temporal and timbral characteristics of the original signal.

**Additive and subtractive synthesis**

*Additive synthesis* is based on the Fourier Series approximation, by which any periodic signal can be decomposed into a (potentially) infinite sum of sinusoidal signals.

In practice it means adding a finite number of sinusoidal signals each with their corresponding amplitude [41]. As many oscillators as signals are needed, and the more signals used when synthesizing, the better the approximation to the original signal will be.

Additive synthesis might be an interesting approach for generating the wind-noise part of the masker, although it would require a significant number of oscillators/-source signals to create its broad spectrum. Because of this, a technique such as subtractive synthesis might be more apt.

*Subtractive synthesis* works on the opposite principle as additive synthesis. In this scenario, the starting point is a signal with rich spectral density such as white or pink noise. The noise is then filtered using any combination of low-pass, high-pass or band-pass filters to arrive at the desired result.

The basic principle of subtractive synthesis is to use a signal with rich spectral content to filter it with a time-varying resonant filter [21]. Figure 2.11 shows a typical block diagram of subtractive synthesis from the late 1970s. It includes two oscillators, a filter, and two envelope generators (ADSR[3]).

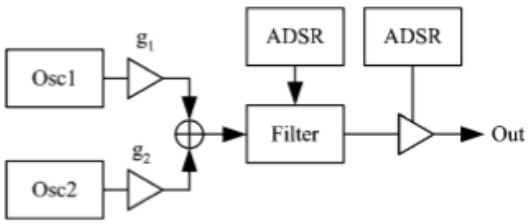



Figure 2.11: Typical block diagram of subtractive synthesis, from [21]

**Combination**

Dubnov et al. [8] used *wavelet-tree learning* to decompose a signal into a wavelet coefficient tree, treating the input sound texture as a sample of a stochastic process. The multiscale wavelet tree signal and structure representation is then resampled by reorganizing the order of the paths down the tree structure. Each path is then used to resynthesize a part of the signal by the inverse wavelet transform. While they generally obtained good results, when using their algorithm to synthesize splashing water on shore they observed that it created repetitions of short splashes that were not apparent in the source, creating the sensation of more "nervous" splashing. It seems clear that a "nervous" repetition of the leaves noise might be perceived as unpleasant and disturbing by the end user, which seems to discourage from using this method.

[3]Attack, Decay, Sustain, Release

# Chapter 3

# Synthesizing the Masker

The previously described physical system developed by the HS-OWL was the limiting factor when deciding the type of masker to implement. Using plastic, a metal frame, and an exciter, a leaf-like structure was designed that would serve as the visual element in the offices that chose to use the system. The exciter is used to emit a very low frequency (in the range of 10 Hz) that physically moves the leaves, giving the sensation that they are rattling in the wind. This structure and mechanism serves to establish a sensory connection as the source of the auditory masker. To simulate this physical phenomena, two separate maskers were developed: a leaf noise and a wind noise. The combination of the two produces the final masker. These two separate maskers also serve to cover the entire frequency spectrum, with the wind mostly serving to mask the low frequencies, and the leaves the middle and high frequencies.

The masker was implemented in MATLAB chosen over other options like visual programming languages such as PureData[1] due to the preexisting familiarity with it, as well as the ability to create the graphical interfaces that would be used to design the listening test that evaluated the masker.

To serve as the basis for what is to be accomplished, the key objectives described in section 1.3 (page 16) are listed below:

- It must be perceived as real, and contain elements of both wind noise and leaves noise to match the physical structure
- It must be accepted by the end users (i.e. not found to be unpleasant or disturbing)
- It must be of an arbitrary duration and not present significant changes in spectral or temporal characteristics

[1] `http://puredata.info/`

## 3.1 Design of the Leaf-Noise Masker

In an extensive study of different signal synthesis methods, Misra and Cook [25] determine that among other methods such as LPC or stochastic models, granular synthesis, is especially apt for creating textures and soundscapes (i.e. not pitched sounds). Due to the relatively complex nature of the sound of leaves, and in order to obtain as real of a result as possible, granular synthesis was used to perform the synthesis. Prior to the granulation, and because of the relatively high density of high frequency information that the leaf sound contains, the source was pitched down. The entire block diagram of the leaf noise synthesis can be seen in figure 3.1.

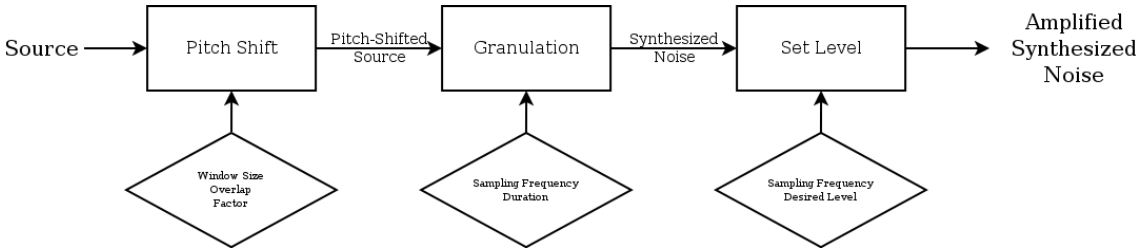



Figure 3.1: Block diagram of leaf synthesis algorithm

### Source Analysis

In order to use granular synthesis, one must have a source file from which to take the grains. In this case, a cleanly-recorded leaf sound was needed. A 43-second sample was found on Freesound[2], which was then inspected for a snippet that contained only the desired leaf sounds. This file, which will now be referred to as "source.wav"[3], contains the audio from 0.215 seconds until 7.540 seconds from this source. Its waveform and spectrum can be seen in figure 3.2. While the sound of a bush being rustled, rather than leaves in a tree, it provides sufficient source material to obtain the desired leaf sound.

### 3.1.1 Pitch Shifting

As the spectrum in figure 3.2 illustrates, the obtained leaf sound source has a very high concentration of high frequencies, which over long exposure times as might be the case with the masker can become very unpleasant to listen to. One option to remove these high frequencies would be to use a low-pass filter. However, this method leads to a significant loss of information of the original signal, which results in a very unnatural-sounding result.

[2] `https://www.freesound.org/people/duckduckpony/sounds/204030/`
[3] The "source.wav" file can be found in the attached CD.

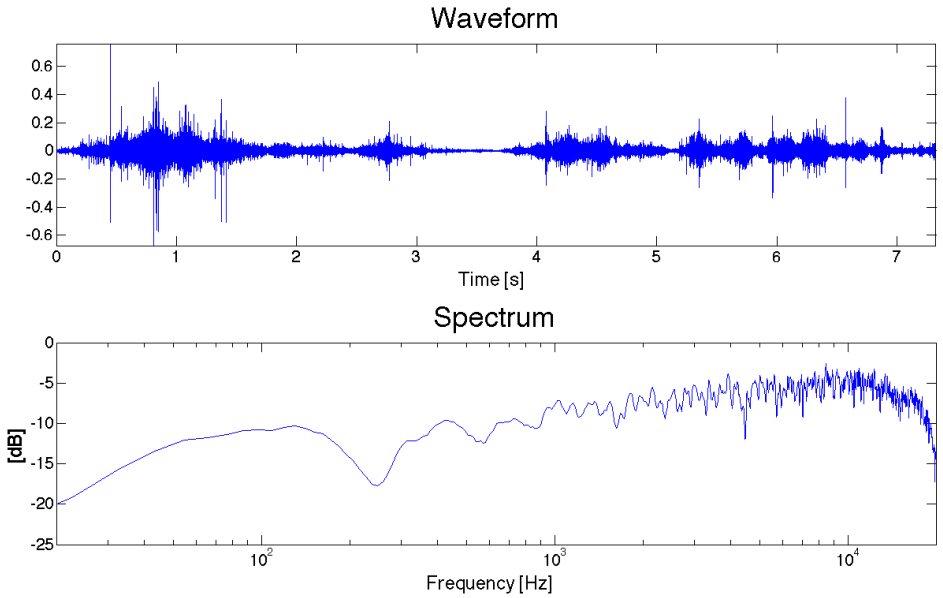



Figure 3.2: Waveform and spectrum for the original source

Instead, pitch-shifting was used to obtain a more pleasant-sounding, acceptable signal. There were two possible options as to when to perform pitch-shift: 1) on the original source, prior to the synthesis or 2) after the synthesis. Figures 3.3 and 3.4 show the spectrums of the original and pitch-shifted source, and the synthesized noises. Though visually similar, performing the pitch-shifting on the final synthesized signal resulted in a highly disturbing sibilance. Therefore, it was opted to perform the pitch-shifting operation prior to the synthesis.

How much to pitch-shift the source signal was determined by trial and error. Taking into account that the high-frequency (from 10 to 12 kHz) content of the leaves sound was not important as human speech does not extend into that range, it was opted to lower the pitch of the leaves by the maximum amount possible without adding excessive artifacts. This amount was found to be a 15% shift toward the lower frequencies.

Though it is true that the pitch-shifting alters the resulting sound, which therefore becomes less "real", given that the leaf-sound is still distinguishable, it was considered an appropriate sacrifice. After all, no matter how "real" the resulting synthesized noise is, if it is perceived as disturbing it won't be accepted by the users.

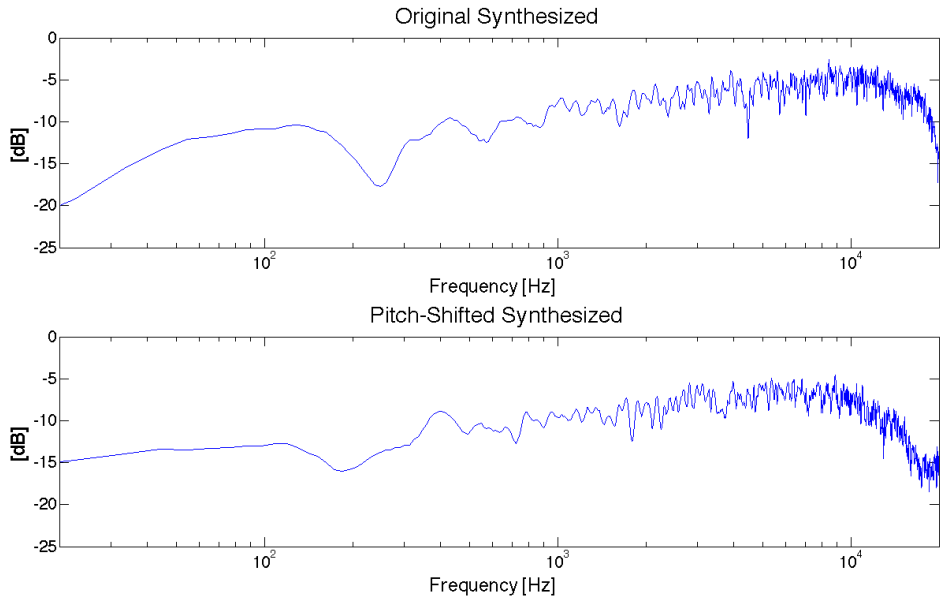



Figure 3.3: Spectrum of the original and pitch-shifted source

### 3.1.2 Granular Synthesis

After the original source sound has been successfully pitch-shifted, the next step is to use granular synthesis to obtain a signal of arbitrary length.

As described in section 2.5 (page 31), granular synthesis consists of using a source sound, and splitting it into small fragments of audio, called grains, which are used to synthesize a new signal. Either short or long grains can be used. When using short grains, the result is a very pulse-like signal, whereas longer grains allow to maintain some of the temporal and timbral characteristics of the original signal.

The general structure of a granulation-based granular synthesis approach is roughly the same: first the grains must be somehow obtained from the source, normally via some kind of segmentation, and then these must be pieced back together in a semi-aleatory sequence to produce a new signal. Figure 3.5 shows the most basic implementation of granular synthesis. In it we see the original signal split into six different grains, which are then reshuffled to form a new, synthesized signal. For simplicity, this is a particularly simple example, as each grain is used only once to form a synthesized signal of the same duration. In actuality, each grain can be used more than once or multiple grains can be created, with overlaps, to create a signal of an arbitrary length.

In the granulation algorithm there are three key parameters that ultimately decide the final structure, and therefore sound, of the synthesized signal. First there is the *number of events*, which refers to the number of grains that are ultimately

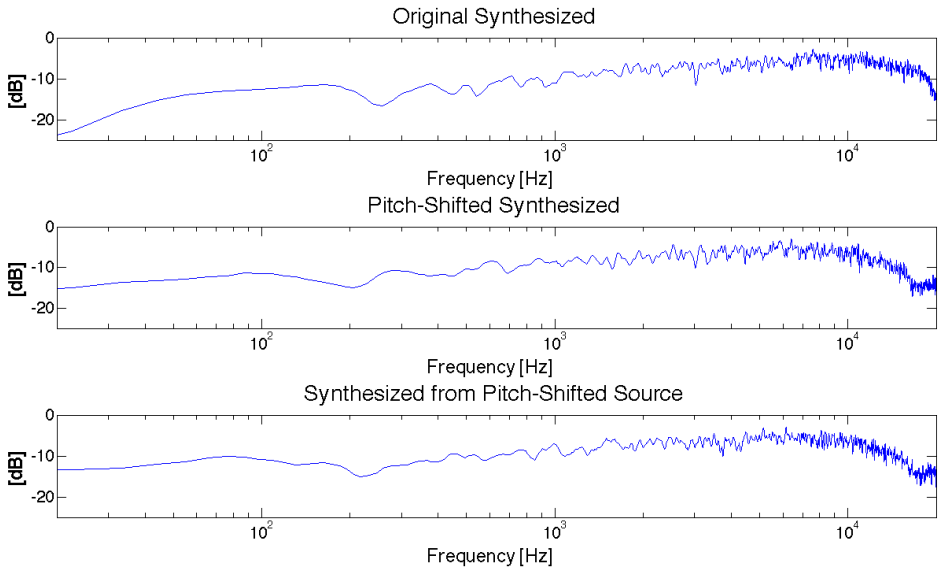



Figure 3.4: Spectrum of the synthesized leaf noise, pitch-shifted synthesized leaf noise, and synthesized leaf noise from pitch-shifted source

created from the source - with no limits as to overlaps -, and that will be used to generate the resulting signal. This parameter establishes the "density" of the signal: a low number of events will result in a very sparse signal, with a lot of space between the grains, whereas a very high number of events will result in nearly indistinguishable noise. The second and third parameters establish the length of the grains used: one sets the minimum grain length, and the other the maximum one. As previously mentioned, longer grains allow to better capture the longer-term temporal characteristics of the source signal. The actual grain length will be a random value in between the minimum and maximum grain length to allow for a more naturally random result, and obviously having little difference between the maximum and minimum grain length will lead to more uniform grains. Examples of three different grains can be seen in figure 3.6.

For the scope of this thesis a relatively simple algorithm was used, using relatively long grains that are arbitrary slices of the source signal. Since the goal of the synthesized signal is to serve as a masker, and the leaves will be used to mask the higher frequency components of speech, a relatively high-density signal will be needed. While this provides a more "constant" flutter of leaves that will most likely not be the case in the real world, it should result in a more consistently present noise. This should reduce the changing-state characteristics that have proven to be disruptive and lead to the Irrelevant Speech Effect as discussed in section 1.1.

Though a relatively constant noise is desired, it is important to maintain the sonic characteristic of the leaves fluttering in the wind. Therefore, a relatively long grain

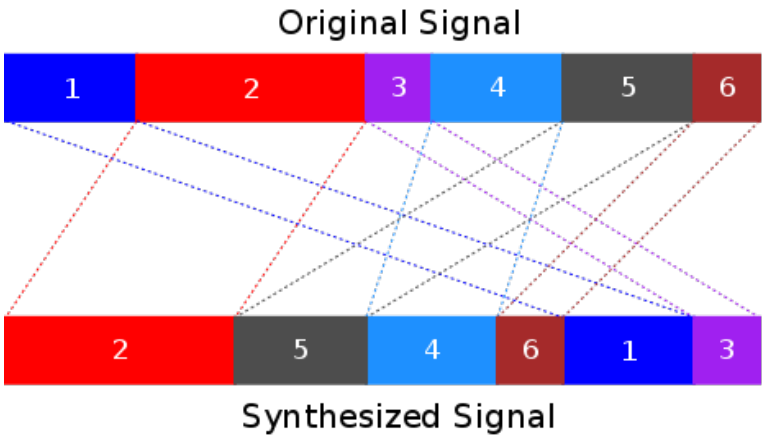



Figure 3.5: Example of simple granular synthesis (without overlap, and using each grain once)

was used, with a minimum length of 150 milliseconds and a maximum length of 300 milliseconds. This allowed to capture some of the long-term fluttering patter, while still maintaining a relatively constant rustling. By not using an excessively long grain, it is hoped that the changing-state characteristics are reduced so as to not be perceived as distracting.

An important thing to note in the implemented algorithm, is that it is iterative depending on the length of the source material. Each iteration of the algorithm provides exactly as much sound as the original source; if the source is seven seconds of audio, and 14 seconds are desired, it will run twice and then join the two streams together. If the desired amount of synthesized signal is (as will often be the case) not an exact multiple of the amount of source, the output signal will be truncated. For example, if 16 seconds of synthesized signal are desired with that same seven-second source, the algorithm will perform three iterations for a total of 21 seconds of signal, and then take from it only the first 16. While this method may be more computationally expensive - which is not a constraint in this case as the algorithm does not have to run in real time -, particularly with longer source audio file, it also allows for a more refined control of the above mentioned parameters. Particularly the number of events stops being dependent on the length of the desired output, which can be key when generating long streams of synthesized audio. One must only think of how "dense" of an output is required. Though establishing a relationship between the number of events and the length of the output with respect to the source was attempted, it was not successfully achieved.

Figure 3.7 shows the waveform and spectrum of a 10-second example of a synthesized leaf-noise. As can be observed, though of a higher density than the original source (figure 3.2), the temporal characteristics, marked by the peaks are relatively well preserved.

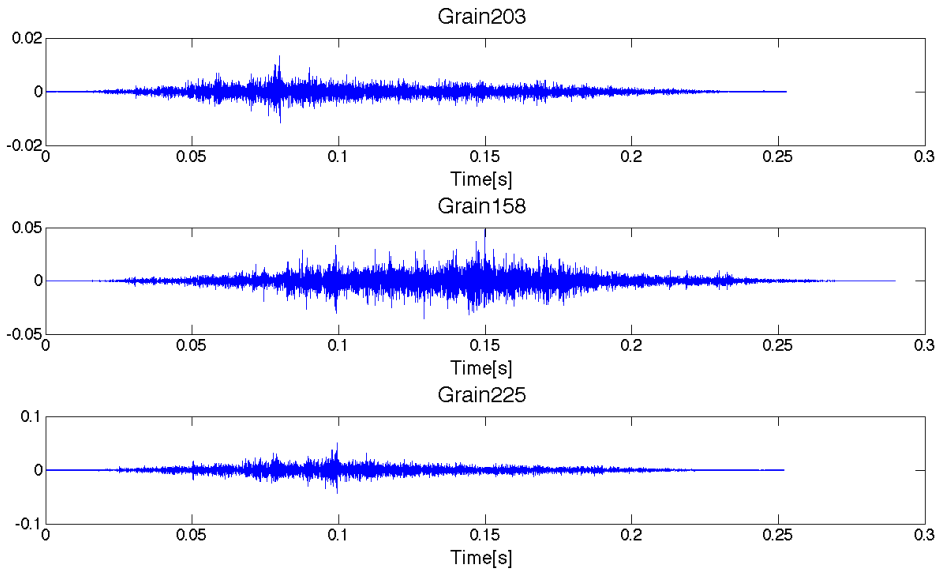



Figure 3.6: Examples of three random grains from the granular synthesis

## 3.2 Design of the Wind-Noise Masker

Wind-like sounds have previously been successfully synthesized using by filtering white noise [48]. Though Verron demonstrates a much more complex method of synthesizing wind in his PhD thesis [47], the spectral characteristics he illustrates in his diagrams (one of which can be seen in figure 3.8), the spectral characteristics of white noise provide the basis from which filter from.

As described in section 2.5.2 (page 31), subtractive synthesis starts with a spectrally rich signal that is then filtered using low-pass, high-pass or band-pass filters to obtain a desired result. When necessary — normally for musical synthesis —, envelope generators can also be used to give a temporal characteristic to the synthesized signal.

Therefore, subtractive synthesis was used to create the wind masker, using a low-pass and a high-pass filter, each with variable cutoff frequencies. Figure 3.9 illustrates the block diagram of the employed algorithm, which will be explained more in detail below.

### Algorithm

The algorithm used for the wind noise synthesis can be seen in figure 3.9. It is composed of five different blocks, each with their own relevant parameters, which will be described in detail in this section.

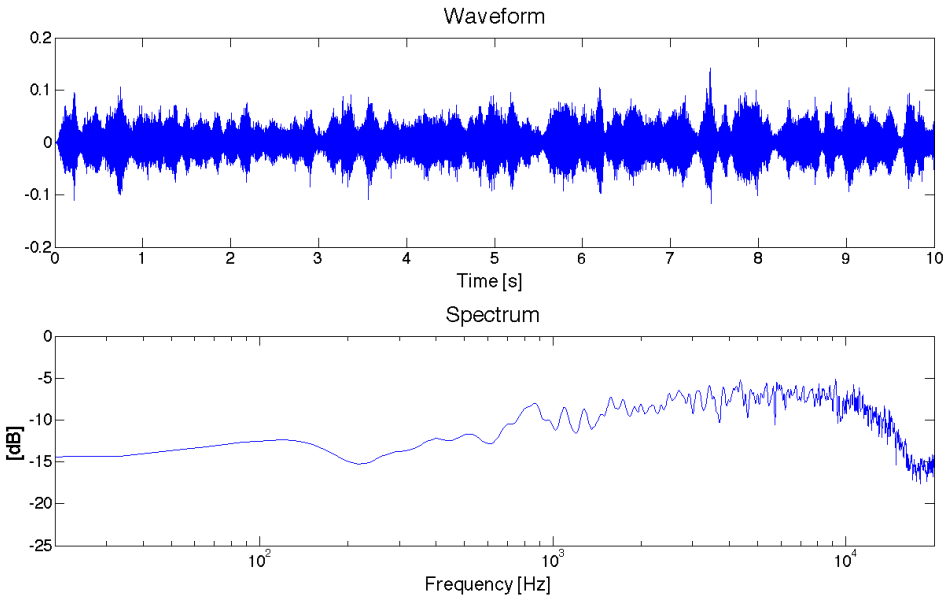



Figure 3.7: Spectrum and waveform of a 10-second synthesized leaf-masker

### Generator

Generally, signals with some periodicity and that are rich in harmonics, such as sawtooth or square signals are used for musical applications of subtractive synthesis. To create a relatively random signal, however, such as the wind noise, it is more appropriate to use a starting signal as a source that contains little harmonic structure. White or pink noise would be apt starting points, but white noise was chosen due to its equal distribution of frequencies. Since no specific temporal envelope is known, contrary to what might be the case in the synthesis of musical instruments, none will be used.

The generator, then, outputs a white noise of the given duration. Figure 3.10 shows the spectrum of the resulting white noise signal. As can be observed it has a fairly even frequency distribution, making it ideal as a source.

### Frame

The output of the generator is then split into frames. Framing the signal is essentially splitting it into smaller parts that can each be processed individually at a later point in time. Though theoretically an unnecessary step, creating frames allows to vary the cutoff frequencies of the low-pass and high-pass filters used in the following steps.

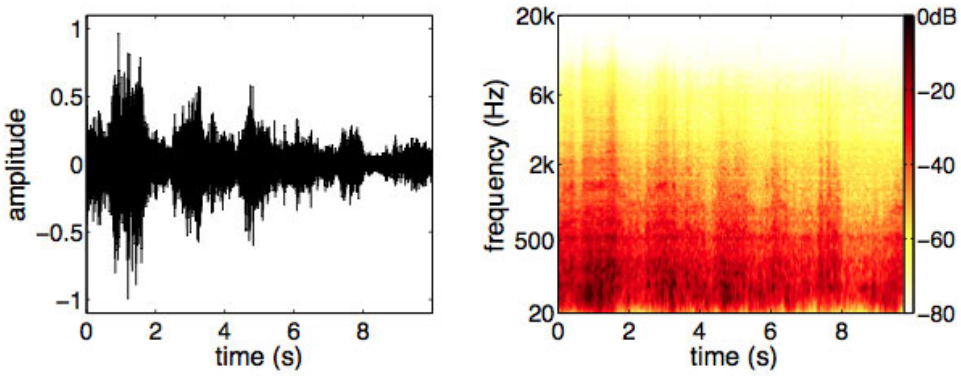



Figure 3.8: Analysis of wind sound, from [47]

### Low-Pass and High-Pass Filters

The core of the subtractive synthesis algorithm is filtering the white noise signal to obtain the desired sound. In this case, that means using both a low-pass and a high-pass filter in order to shape the white noise into something that sounds more like wind. Each of these filters is a high-order (8) Butterworth filter. This kind of filter was chosen due to their flat magnitude response in the pass-band, and the relatively steep slope of $-20n$ dB/decade, where $n$ is the filter order, in this case 8.

When filtering, it is necessary to establish a cutoff frequency. Looking at the analysis of a wind sound that Verron performed (3.8) shows that most of the energy is concentrated in the lower frequency range, below 1000 Hz. This overlaps nicely with the synthesized leaf-noise, which has most of its spectral energy above this point. Therefore, the cutoff frequency of the low-pass filter must be established near 1000 Hz. That same diagram also shows that the energy isn't concentrated in the very low frequencies. Rather, most of it seems to be concentrated from around 100-300 Hz. In addition, the human voice doesn't extend to such a low frequency range, and therefore having the masker act then wouldn't be effective. It is also important to remove these deep bass frequencies as they produce a rumbling effect, which could be disturbing.

Based on the above, and considering the variation in the cutoff frequency that is described below, the cutoff frequencies of the low-pass and high-pass filters were estimated through trial and error to obtain a relatively realistic-sounding wind noise. A cutoff frequency of 880 Hz was selected for the low-pass filter, and a cutoff frequency of 250 was selected for the high-pass filter.

### Variable Cutoff Frequency

One of the main characteristics of a wind sound is that it is relatively dynamic.

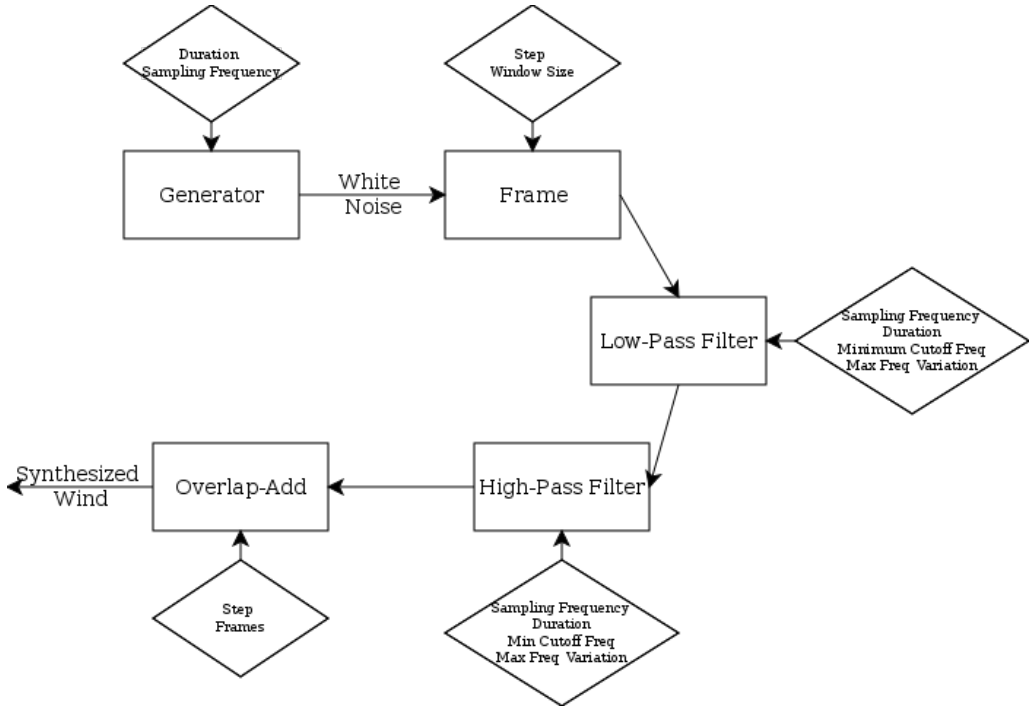



Figure 3.9: Wind noise synthesis algorithm block diagram

Maintaining a constant cut-off frequency isn't enough to produce the effect, and therefore some variation is needed. Since the white noise signal has previously been split into frames, each one of these frames will be filtered using a different cutoff frequency within a specified range.

The variation in the cutoff frequency was again obtained through trial and error, taking into account the parameters described above. That is, it was sought that the wind-noise extended from around 100 Hz up to at least 1000 Hz. Ultimately, it was decided that the variation should be half of the value of the cutoff frequency, in order to produce a more dynamic-sounding wind. Therefore a variation of 440 Hz was established for the low-pass filter, and a variation of 125 Hz was established for the high-pass filter.

Figure 3.11 shows the results of filtering the white noise with the variable-cutoff frequency high-pass and low-pass filters individually. The base cutoff frequency is represented by $f_c$, and the variation by $\Delta f_c$.

### Overlapp-Add

After filtering the white noise frames with the variable-cutoff high-pass and low-pass filters, it is necessary to reassemble them in order to have a complete signal. Since what is obtained is a random signal with little time-dependent information, and in

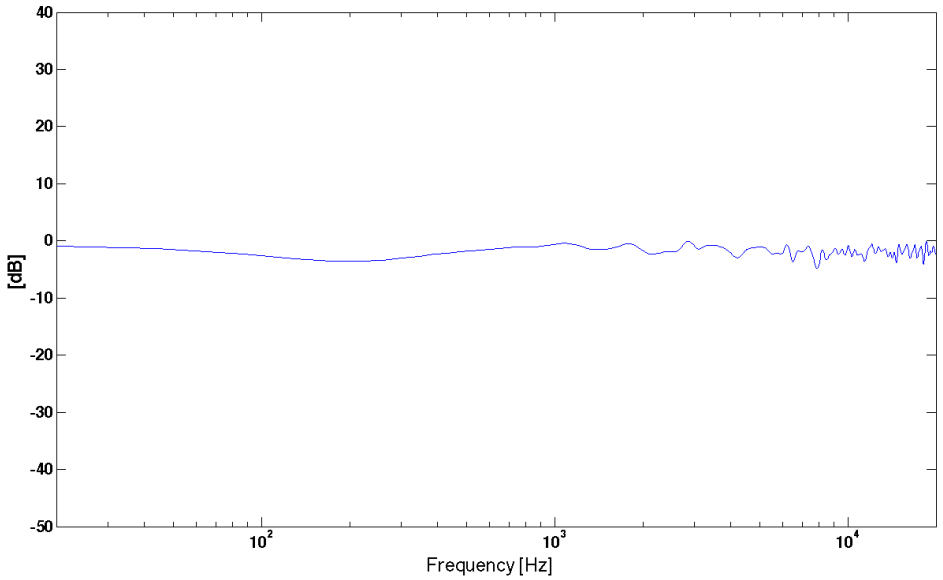



Figure 3.10: White noise spectrum

order to avoid discontinuities that might produce audible clicks, the overlap add method is used to add the frames together.

The output is the final, synthesized wind signal. Figure 3.12 shows the resulting waveform and spectrum of the synthesized wind-noise masker.

## 3.3 Implementation in MATLAB

As has been previously explained, MATLAB was chosen to implement the noise maskers. MATLAB is a "high-level language and interactive environment for numerical computation, visualization, and programming". It's specifications include "[analyzing] data, [developing] algorithms, and [creating] models and applications... [reaching] a solution faster than with spreadsheets or traditional programming languages"[4], making it perfect for the kind of data processing that was performed to synthesize the masker.

The leaf-noise and wind-noise maskers were synthesized separately, and then later combined at different levels to form the final speech masker. The following sections describe the functions written to synthesize the maskers. All of the MATLAB files mentioned can be found in the attached CD.

[4] `http://www.mathworks.de/products/matlab/`

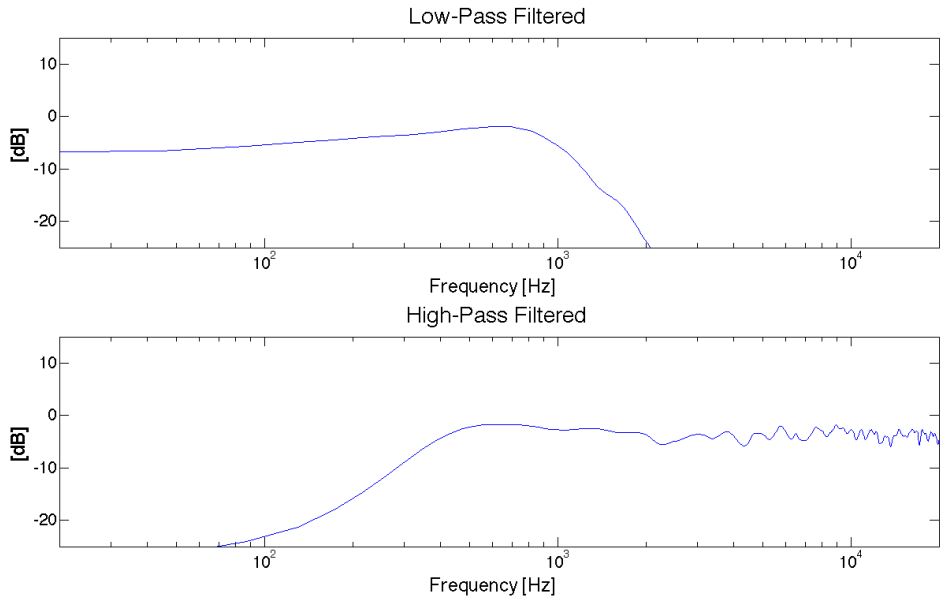



Figure 3.11: Low-Pass and High-Pass filtered white noise spectrums. Low-Pass with $f_c = 880$ and a $\Delta f_c = 440$ and High-Pass with $f_c = 250$ and a $\Delta f_c = 125$

### 3.3.1 Leaf-Noise Masker

The block diagram of the implementation of the leaf noise masker was shown in figure 3.2. As can be seen, it consists of three different blocks: a first block in which the original source sound was pitch-shifted, a second in which granular synthesis was used to create a signal of a given length, and a third in which the appropriate level was set. Each of these blocks corresponds to a different MATLAB function; the pitch-shift and granulation blocks are described below, whereas the level is described in section 3.3.3 as it is also used when implementing the wind-noise masker.

#### Pitch-Shift

The main function of the pitch-shift algorithm is *pitchShift*, whose information can be seen in table 3.1. The pitch-shift code has been adapted from the one provided by Grondin [14] as part of his developed "Guitar Pitch Shifter"[5].

Grondin's algorithm is meant to be used for musical applications. Therefore, the inputs were meant to be the number of semitones of the pitch shift. This was changed to instead be a factor that is directly interpreted as the amount by which to pitch-shift. The bulk of the rest of the code is the same, as it is based on the algorithm Grondin describes, and which is described in detail in appendix A.

[5] `http://www.guitarpitchshifter.com/matlab.html`

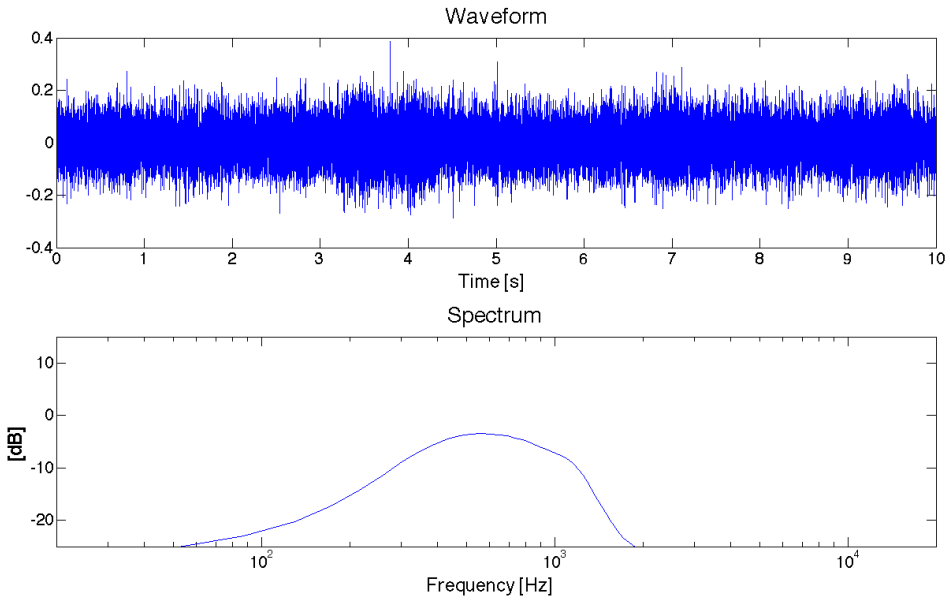



Figure 3.12: Waveform and spectrum of the final synthesized wind noise

The pitch-shift code also requires the use of the functions *frame* and *overlapAdd*, which are described in section 3.3.3.

Table 3.1: MATLAB function *pitchShift* information

| **pitchShift** | |
|---|---|
| **Description** | Pitch-shifts a given source signal by some factor by doubling the length and then resampling |
| **MATLAB Syntaxis** | *pitchShifted = pitchShift(original, winSize, overlap, factor)* |
| **Inputs** | *original*: Original signal to be pitch-shifted<br>*winSize*: Size of the Hanning window<br>*overlap*: Amount of overlap between frames (between 0 and 1)<br>*factor*: Amount by which to the original signal will be pitch-shifted |
| **Outputs** | *pitchShifted*: Resulted pitch-shifted signal |

## Granulation

The granulation block of the algorithm consists of two functions *granulation* and *grainLn*. The former contains most of the necessary code, and is used to perform the granular synthesis to generate the signal of the specified duration. The latter is a function that returns a grain of a specified length from a given source. Tables 3.2 and 3.3 show each functions' respective inputs and outputs.

Marshall's [24] code[6] was used as the starting point for the granular synthesis, with considerable changes. His initial code generated a synthesized signal of the exact length of the source material, which was not sufficient for the requirements of the

[6] `http://www.cs.cf.ac.uk/Dave/Multimedia/Lecture_Examples/Granular.zip`

masker. Therefore, the main change made was the implementation of repeated iterations that concatenated synthesized signals of the same duration as the source until the desired duration was reached.

This method was opted instead generating only one signal of the desired length so as to have more control of the characteristics of the signal. The parameters *nEv*, *minL*, and *maxL* allow for the control of the number of events (i.e. grains) used in the synthesis, as well as the minimum and maximum grain length. As described in section 3.1.2, the length of the grain establishes the temporal characteristics of the synthesized signal, with longer grains allowing for more the long-term temporal characteristics, and shorter grains creating a more impulsive signal. The number of events establishes how dense the synthesized signal will be. A low number of events will result in a sparse signal, and a high number in a very constant one. Since it is difficult to establish a relationship between the total duration of the signal and the number of events, synthesizing in smaller parts allows for a finer control.

These three parameters were established through extensive trial and error, taking into account as described in section 3.1.2 that since the synthesized signal was going to be used as a speech masker, it should be relatively present (i.e. dense), and should preserve most of the temporal characteristics of the original. The number of events was determined to be 500, and the minimum and maximum grain lengths were established to be 150 and 300 milliseconds, respectively.

Since the source signal contained relatively constant leaf noise, the exact selection of the grains was not deemed to be especially relevant. Therefore, the selection of the grains from the source signal was done at random points, for simplicity, allowing for overlap and for the possibility of not necessarily using the entire source signal. This approach was also taken when placing the grains, taking care that they covered the entirety of the output.

Table 3.2: MATLAB function *granulation* information

| | **granulation** |
|---|---|
| **Description** | Creates an output signal of the specified length using granular synthesis and the given source |
| **MATLAB Syntaxis** | *synth = granulation (source, fs, dur)* |
| **Inputs** | *source*: Source signal to be used for the granular synthesis<br>*fs*: Sampling frequency<br>*dur*: Desired duration of the synthesized signal |
| **Outputs** | synth: Synthesized signal of duration dur |

The grains themselves are simply made by taking a segment of a given length from the source file, from a given starting point. The output is then returned to be used in the granulation function.

Equation (3.1) shows the granular synthesis equation, in which the output signal $y$ is a result of adding a *grain* to itself. A random amplitude $A$ is applied to each grain so as to have a more "flowing" effect. Equation (3.2) illustrates how each grain

Table 3.3: MATLAB function *grainLn* information

| grainLn | |
|---|---|
| **Description** | Creates a grain of the specified length and starting point from a source signal |
| **MATLAB Syntaxis** | *grain = grainLn (source, init, L)* |
| **Inputs** | *source*: Signal from which to take the grain<br>*init*: Initial sample for the grain<br>*L*: Length of grain (in samples) |
| **Outputs** | *grain*: Grain of length *L* taken from *source* |

is calculated. A Hanning window is used to smooth the grains, and create fade ins and fade outs.

$$\mathrm{y}[j : j + L] = \mathrm{y}[j : j + L] + A \cdot \mathrm{grain}[i : i + L] \tag{3.1}$$

$$\mathrm{grain}[i : i + L] = x[i : i + L] \cdot \mathrm{w}_{\mathrm{Hanning}}[n] \tag{3.2}$$

### 3.3.2 Wind-Noise Masker

Though the block diagram of the wind-noise masker (figure 3.9, page 44) shows five different blocks, it is generated in only one function, *synth_ wind*. This function has one external call to the function *randomCutoff*, which generates the variable cutoff frequencies that are used for filtering the noise. Table 3.4 shows the inputs and outputs of the *synth_ wind* function.

Table 3.4: MATLAB function *wind_ synth* information

| wind_synth | |
|---|---|
| **Description** | Synthesizes a wind-like noise of a specified duration |
| **MATLAB Syntaxis** | *wind = synth_ wind(dur,fs,step,frameSize)* |
| **Inputs** | *dur*: Desired duration of the wind noise<br>*fs*: Sampling frequency<br>*step*: Space between frames<br>*frameSize*: Size of the frames |
| **Outputs** | *wind*: Synthesized wind noise of duration *dur* |

Below is the wind-synthesis algorithm:

```
Create white noise signal of 1.3 times desired duration (a longer length
    is used because some length is lost during overlap-add)

Define center cutoff frequencies and maximum variation for both the
    high-pass and low-pass filter
```

```
Divide the white noise into frames according to the input size and step
    between frames

Calculate array variable cutoff frequencies for the low-pass and
    high-pass filters

Frame the arrays of variable cutoff frequencies according to the input
    size and step between frames

For each variable cutoff frequency frame
        Calculate the mean of the cutoff frequencies
        Convert the mean to a number between 0 and 1, where 1 is half the
            sampling frequency

For each white noise frame
       Window the frame with a Hanning window
        Low-pass filter the frame with a Butterworth filter of order 8
        High-pass filter the frame with a Butterworth filter of order 8

Overlap-add the filtered frames

Truncate to only the desired length
```

### Variable Cutoff Frequency

Using a variable cutoff frequency creates a sense of dynamism to the wind noise, making it seem more real. This variable cutoff frequency is generated by the function *randomCutoff*, whose inputs and outputs can be seen in table 3.5.

Table 3.5: MATLAB function *randomCutoff* information

| **randomCutoff** | |
|---|---|
| **Description** | Generates an array of cutoff frequencies within a specified range |
| **MATLAB Syntaxis** | *fc = randomCutoff(fs,dur,fMin,fDeltaMax)* |
| **Inputs** | *fs*: Sampling frequency<br>*dur*: Duration of the signal that will be filtered<br>*fMin*: Minimum cutoff frequency<br>*fDeltaMax*: Maximum change of the cutoff frequency |
| **Outputs** | *fc*: Array of $fs \cdot dur$ cutoff frequencies |

The MATLAB code, with comments, for this function is below.

```
factor = 20; % Rate of change factor
fc = randn((fs/factor)*dur,1); % Create random slow variations

fLow = 0.00001; % Cutoff frequency
B = fir1(1000,fLow,'low'); % Hard lowpass
fc = filtfilt(B,1,fc); % Filter
```

```
fc = fc .* (fc>0); % Cut negative values

fc = resample(fc,factor,1); % Resample on sampling frequency

% Avoid filtering artifacts
fc(1:10000) = 0;
fc(end-10000:end) = 0;

% Scale
fc = fc * fDeltaMax / max(fc);
fc = fc + fMin;
```

### 3.3.3 Auxiliary Functions

There were three functions that were used for both the leaf-noise masker and the wind-noise masker: *soundALevel*, *frame*, and *overlapAdd*. The first is used to set a given input signal to a desired A-weighted level. The second splits an input signal into frames, and the third is a function that overlap-adds a series of frames to give an output signal.

#### soundALevel

The function *soundALevel* is used to set an input signal to a given A-weighted level[7]. It is used essentially as an amplifier to set the synthesized noises to normalized levels for testing. To do so, the input signal is first equalized according to the A-weighting curve. The Root Mean Square (RMS) value of the desired A-level value is calculated from the given dB value, according to equation (3.3). A factor is obtained by dividing this value by the RMS value of the original signal, and the original signal is then multiplied by it.

The inputs and outputs of this function can be seen in table 3.6.

$$ALevel_{\text{rms}} = p_{\text{ref}} \cdot 10^{(\frac{ALevel_{\text{dB}}}{20})} \tag{3.3}$$

#### Frame and Overlap-Add

The *frame* and *overlap-add* functions are used to split a signal into frames for processing and then add them back together, respectively. Their respective inputs and outputs can be seen in tables 3.7 and 3.8.

[7] A-weighting is one of the curves defined in the IEC 61672:2003. It is applied to sound measures to compensate for the relative loudness of human hearing.

Table 3.6: MATLAB function *soundALevel* information

| **soundALevel** | |
|---|---|
| **Description** | Sets the level of a given signal to a specified A-weighted value |
| **MATLAB Syntaxis** | *y,= soundALevel(x, fs, Alevel, pref)* |
| **Inputs** | *x*: Input signal<br>*fs*: Sampling frequency<br>*Alevel*: Desired A-weighted level<br>*pref*: Reference pressure |
| **Outputs** | *y*: Output signal with level *Alevel* |

Table 3.7: MATLAB function *frame* information

| **frame** | |
|---|---|
| **Description** | Splits a signal into frames |
| **MATLAB Syntaxis** | *[frames numFrames] = frame(signal,step,frameSize)* |
| **Inputs** | *signal*: Original signal to be split into frames<br>*step*: Space between frames<br>*frameSize*: Size of the desired frames |
| **Outputs** | *frames*: Matrix of *numFrames* by *frameSize* elements containing the resulting frames<br>*numFrames*: Number of resulting frames |

Figure 3.13 illustrates the concepts of framing and overlap-adding, and is a visual representation of how the functions are each coded. Generally a frame size of 1024 was used, with a step of 256 for a 75% overlap.

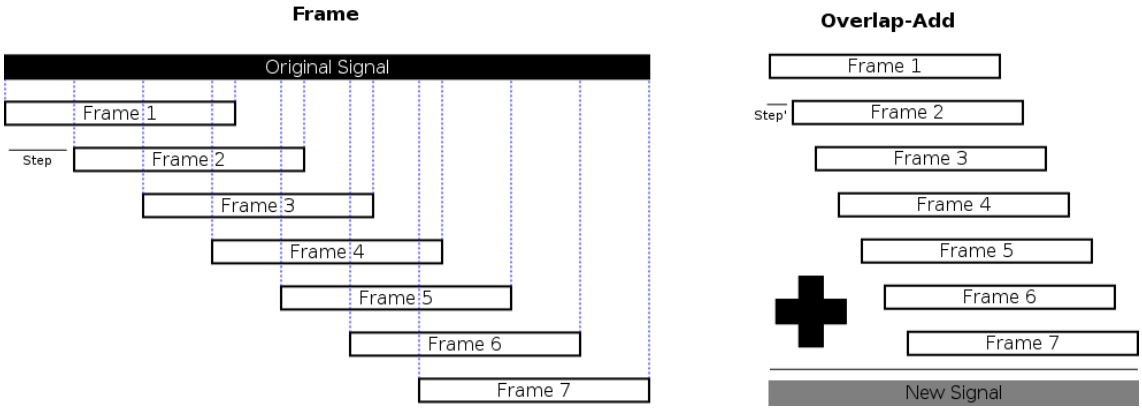



Figure 3.13: Diagram illustrating the concepts of framing (left) and overlap-adding (right)

## 3.4 Obtaining the Final Speech Masker

This chapter has described how each of the individual maskers have been synthesized. The leaf-noise was generated using granular synthesis, with a pitch-shift done prior to the synthesis. Subtractive synthesis using variable cutoff low-pass and high-pass filters was used to synthesize the wind-noise masker.

Table 3.8: MATLAB function *overlapAdd* information

| **overlapAdd** | |
|---|---|
| **Description** | Overlap-adds a series of frames into a complete signal |
| **MATLAB Syntaxis** | *newSignal = overlapAdd (frames,step)* |
| **Inputs** | *frames*: Matrix of frames to be overlap-added<br>*step*: Space between frames |
| **Outputs** | *newSignal*: Signal resulting from the overlap-add of the *frames* |

The objective is to have only one masker that will serve to mask speech and that contains both of the elements of the individually synthesized maskers. Therefore, these must be combined. This is done simply by adding them together, sample for sample. It is important to have leveled them first so as to add the same "amount" of each, or, alternatively a known difference. As will be described in detail in chapter 4, five variations of the final masker were generated, corresponding to different level differences between the leaf-noise masker and the wind-noise masker. These five maskers are described in table 3.9.

Table 3.9: Description of test labels

| **Label** | **Masker** |
|---|---|
| *white* | White noise |
| *pink* | Pink noise |
| *masker_0* | Synthesized masker with 0 dB(A) difference between leaves noise and wind noise |
| *masker_p2* | Synthesized masker with +2 dB(A) difference between leaves noise and wind noise |
| *masker_p4* | Synthesized masker with +4 dB(A) difference between leaves noise and wind noise |
| *masker_m2* | Synthesized masker with -2 dB(A) difference between leaves noise and wind noise |
| *masker_m4* | Synthesized masker with -4 dB(A) difference between leaves noise and wind noise |

Their spectrums can be seen in figure 3.14. Each of these five variations were obtained at three different levels. This total of 21 different noise maskers will then be evaluated to test how much they affect speech intelligibility, i.e. how well they mask speech.

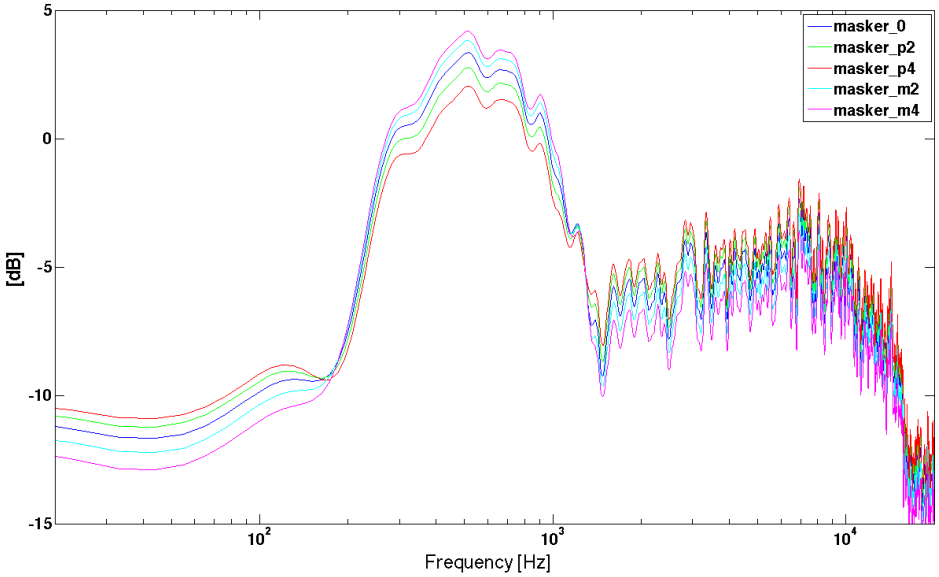



Figure 3.14: Spectrums of the five different synthesized maskers

# Chapter 4

# Evaluating the Masker

After the masking noise had been synthesized as described in chapter 3, the next step was to evaluate it to determine how well it fit the specifications. The main goal of the masking noise is to reduce speech intelligibility so as to reduce speech's effect on performance and concentration. Therefore, a test was designed and carried out to compare how well it did so in comparison to two reference signals: white noise and pink noise. To measure the subjective user acceptability, a questionnaire was be issued within the test.

Other than to examine how well the masker works, the designed speech intelligibility test served to establish under which conditions the masker works best. That is, the test determined the optimal level difference between the synthesized wind and leaves noises as well as how much louder it should be compared to the disruptive speech signal.

## 4.1 Participants

12 participants (11 male and one female), all students or employees of the Erich-Thienhaus Institute of the HfM, took park individually in the voluntary listening test. The ages of the participants ranged from 20 to 36 (mean age = 26.1, standard deviation = 4.2). In exchange for their participation, the volunteers we offered chocolates and/or sweets that were available during the course of the test. So that language was not a potential bias, all test subjects were native German speakers.

Participants were not necessarily told what was being investigated, though some had prior knowledge due to a personal relationship with me. However, possible prior knowledge or not as to the purpose of the test was not considered of importance.

Though not of special relevance due to the nature of the evaluation, the names of the test subjects were not collected. The only personal information that was requested were the subject's sex, age, and occupation, so as to have an accurate description of the sample group.

## 4.2 Apparatus

The listening test was carried out in a room equipped with a 5.1 system of Musikelectronic Geithain RL901K[1] reference loudspeakers, though for the purposes of the test only two (emitting a mono signal from the left and right channels) were used.

Participants sat at a desk three meters from the loudspeakers, at a 60° angle from each, and carried out the test on a Toshiba M10[2] laptop connected through USB 2.0 to an Avid MBox[3] that was connected to the loudspeaker system. The test was generated through a graphical interface created in MATLAB, and that same program collected the test results.

The audio files used in the experiment (see section 4.3 for how they were used) were .WAV files sampled at 44.1 kHz. The noises were generated with MATLAB or Audacity, and the speech samples were obtained from Maximilian Schmitt, who had previously recorded them within the scope of the Private Workspace project. There were a total of 150 speech signals (i.e. words) – 50 words, with three intonation variations of each –, and 21 noise signals.

## 4.3 Procedure

### 4.3.1 Speech Intelligibility Test

In order to evaluate how well the synthesized noise masks speech, a modified Oldenburger Satztest (OSLA) was used. The OSLA is an audiometric test to determine the speech intelligibility threshold in quiet and with a noise signal [19]. It's original purpose is to determine a possible hearing impairment. In this case, the modified version of this test was used to determine how adept the developed noise masker is at impairing speech intelligibility.

[1] http://www.me-geithain.de/index.php/en/studio/products/active-loudspeaker/rl901k

[2] http://www.toshiba.co.uk/discontinued-products/tecra-m10-10i/

[3] http://www.avid.com/US/products/mbox

Table 4.1: The 50 words used for the modified Oldenburger Satztest

| Name | Verb | Numeral | Adjective | Noun |
|---|---|---|---|---|
| Britta | bekommt | zwei | alte | Autos |
| Doris | gewann | drei | große | Bilder |
| Kerstin | gibt | vier | grüne | Blumen |
| Nina | hat | fünf | kleine | Dosen |
| Peter | kauft | sieben | nasse | Messer |
| Stefan | malt | acht | rote | Ringe |
| Tanja | nahm | neun | schöne | Schuhe |
| Thomas | schenkt | elf | schwere | Sessel |
| Ulrich | sieht | zwölf | teure | Steine |
| Wolfgang | verleiht | achtzehn | weiße | Tassen |

The Oldenburger Satztest presents participants with a list of 50 words[4], divided into five columns depending on whether they are names, verbs, numerals, adjectives, or nouns (10 words per column). The frequency of the phonemes in the word lists correspond to their frequency in the German language. During each trial a word was chosen at random from each column to form a sentence of the form: name-verb-numeral-adjective-noun. The subjects are then asked to select the words they have heard, or believe to have heard, (one per column) to form a complete sentence. These 50 words can be seen in table 4.1.

After submitting their personal information (sex, age, occupation), the test subjects were presented with a screen of directions (in English) that gave directions as to their task. If any needed clarification, it was given. From here they were given the choice to practice the exercise they would be asked to complete or to begin the test. Depending on their choice further instructions were provided. During the practice round the tests subjects were informed that there would be no constraints in terms of time or repetitions, and when beginning the testing phase they were informed that they would be repeating 5-sentence sets 42 times under different conditions, with the possibility of a break in between tests if needed. In both cases it was made clear that they would not be receiving any feedback about their performance, and that they would not have the possibility of correcting any mistakes. Before beginning each of the individual 42 tests the participants were notified of what test number they were about to begin. A similar screen was presented upon finishing each of the tests.

The subjects began the test by clicking a large button with the text "Go". From this moment they were presented with a graphical interface consisting of 56 buttons (figure 4.1): one button for each word (50 in total), one button with a question mark ("?") per column to be used in the event that no word was made out, and a button

[4]Though this thesis is written in English, German words were used since the test subjects had a higher German language proficiency

labeled "Next" that served to submit the results. Simultaneously they were exposed to noise signal, which faded in linearly for 1.5 seconds, and one second after the fade in the speech began. Approximately one second after the speech ended, the noise linearly faded out for 1.5 seconds. The speech signal was generated randomly each time, choosing one word from each column, and one of the three available variations of each word. The words were separated by a space of one third of a second of silence.


Figure 4.1: Example of the graphical interface for the speech intelligibility test

It was the task of the test subjects to click on one word from each column, corresponding to the words they perceived to have heard. If they could not make the word out, they were expected to click the question mark button. After the test subjects clicked on a button in a column the rest of the buttons in that column were disabled, preventing them from changing their answer, and the text in the selected one turned blue. Once one word from each column was selected, they were asked to submit their answers by clicking "Next". Depending on how many trials had been done, the "Next" button would either present another "Go" button to begin a new iteration (if less than five had been completed), or present the subjective test described in section 4.3.2 (if five had been completed).

The 42 tests were split amongst 21 tests used with male speech and 21 tests with female speech. The speech was always leveled to 60 dB(A) using the function *soundALevel* described in section 3.3.3 (51). The noise exposures were as follows (in parentheses the "test label" is indicated):

1. Exposure to white noise

2. Exposure to pink noise
3. Exposure to masker, with the following variations:
   (a) *Masker_0*: Masker with 0 dB(A) level difference between the leaves noise and wind noise
   (b) *Masker_p2*: Masker with +2 dB(A) level difference between the leaves noise and wind noise
   (c) *Masker_p4*: Masker with +4 dB(A) level difference between the leaves noise and wind noise
   (d) *Masker_m2*: Masker with -2 dB(A) level difference between the leaves noise and wind noise
   (e) *Masker_m4*: Masker with -4 dB(A) level difference between the leaves noise and wind noise

All of the above noises were leveled to be +10 dB(A), +13 dB(A), and +16 dB(A) (leveled to 70 dB(A), 73 dB(A), and 76 dB(A) using *soundALevel*), higher than the speech. The actual emitted equivalent levels ($LA_{eq}$) for the speech and maskers can be seen in table 4.2.

A test situation consisted of the emission of a speech signal along with one of the above noises at one of the aforementioned levels. The order of the tests was generated randomly each time. During the practice situation the participants were only exposed to the speech signal. The white and pink noise signals serve as references by which to measure the synthesized masker against. Each test situation was completed only once, with the entire test lasting between 60 and 75 minutes.

The MATLAB interface through which the test was created collected whether each selected word was correct or not by means of a "0" (incorrect) or a "1" (correct). An example of the test screen for this interface can be seen in figure 4.1, for the full structure of the interface see appendix B.

### 4.3.2 Subjective Evaluation

Along with participating in the speech intelligibility test, the test subjects also answered three subjective questions about each test situation. Upon the completion of each test, they were presented with a screen of three drop down menus (see figure 4.2) that asked them to assess the "realness" and "pleasantness" on a scale from 0 to 5, with 0 being the least and 5 being the most. The subjects were also asked to rate their concentration during the test using that same scale.

The goal with this evaluation was to have a qualitative assessment of the perceived realism of the masker. The masker will be used in conduction to a physical system

Table 4.2: A-weighted equivalent levels of the signals emitted during the listening test

| **Signal** | $LA_{eq}$ **Alone [dB]** | $LA_{eq}$ **Signal + Male Speech [dB]** | $LA_{eq}$ **Signal + Female Speech [dB]** |
|---|---|---|---|
| Male Speech | 58.3 | - | - |
| Female Speech | 59.7 | - | - |
| White_70 | 66.7 | 67.0 | 67.1 |
| White_73 | 69.5 | 69.8 | 69.8 |
| White_76 | 72.4 | 72.6 | 72.7 |
| Pink_70 | 68.8 | 69.3 | 69.0 |
| Pink_73 | 71.9 | 72.1 | 71.9 |
| Pink_76 | 75.0 | 74.9 | 75.1 |
| Masker_0_70 | 69.4 | 69.5 | 70.0 |
| Masker_0_73 | 72.2 | 72.5 | 72.8 |
| Masker_0_76 | 75.2 | 75.5 | 75.8 |
| Masker_p2_70 | 68.6 | 68.9 | 69.4 |
| Masker_p2_73 | 71.8 | 72.0 | 72.5 |
| Masker_p2_76 | 74.6 | 74.7 | 75.5 |
| Masker_p4_70 | 68.2 | 68.5 | 69.8 |
| Masker_p4_73 | 71.0 | 71.2 | 72.6 |
| Masker_p4_76 | 74.1 | 74.2 | 75.7 |
| Masker_m2_70 | 69.7 | 70.1 | 70.1 |
| Masker_m2_73 | 72.8 | 72.9 | 73.1 |
| Masker_m2_76 | 75.8 | 75.8 | 76.7 |
| Masker_m4_70 | 70.3 | 70.4 | 70.3 |
| Masker_m4_73 | 73.4 | 73.4 | 73.6 |
| Masker_m4_76 | 76.6 | 76.2 | 76.7 |

of moving "leaves" developed by the HS-OWL, and therefore it is important that the synthesized masker is at least perceived as somewhat real. The perceived pleasantness is also important to take into consideration, as if the masker is perceived to be too much of a nuisance its use will also have to be questioned, regardless of the results obtained in the quantitative test. Lastly the perceived concentration gave an idea of how distracted the subject might have been, whether through the noise itself or other factors.

The results of this this test were stored alongside the results from the speech intelligibility test. The interface can be seen in figure 4.2.

## 4.4 Test Results

The description for the test labels used in the charts and graphs that follow in this section can be seen in table 3.9 (page 53), reproduced below. This notation will be used throughout this section.

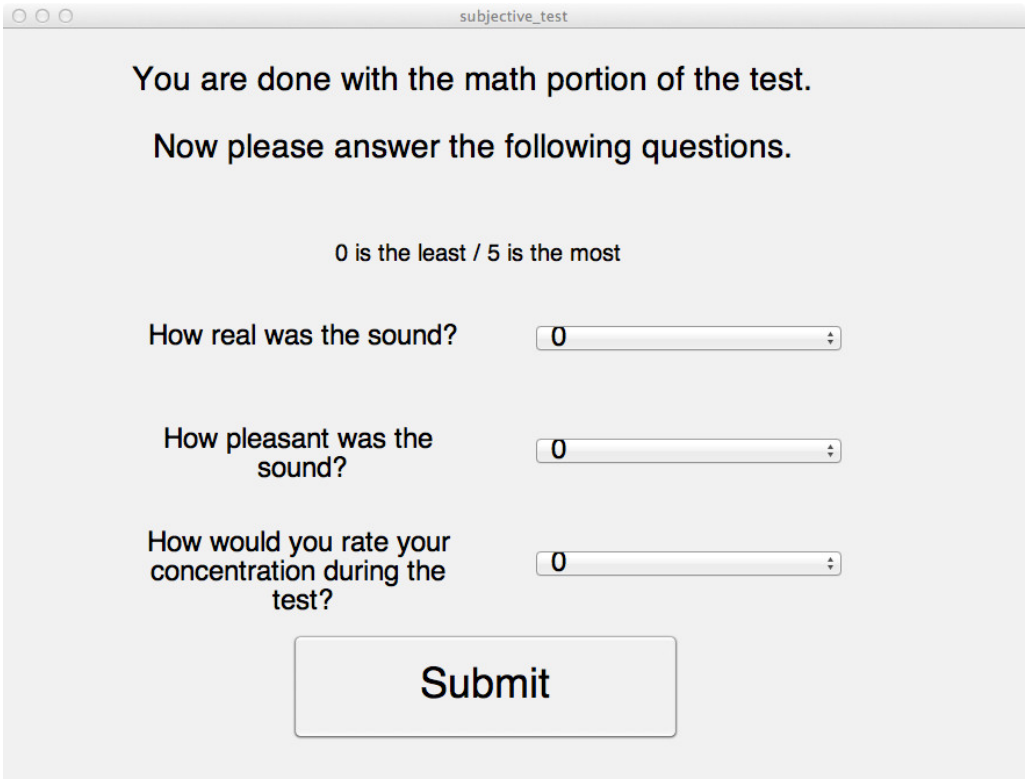



Figure 4.2: Example of the graphical interface for the subjective evaluation

| Label | Masker |
|---|---|
| *white* | White noise |
| *pink* | Pink noise |
| *masker_0* | Synthesized masker with 0 dB(A) difference between leaves noise and wind noise |
| *masker_p2* | Synthesized masker with +2 dB(A) difference between leaves noise and wind noise |
| *masker_p4* | Synthesized masker with +4 dB(A) difference between leaves noise and wind noise |
| *masker_m2* | Synthesized masker with -2 dB(A) difference between leaves noise and wind noise |
| *masker_m4* | Synthesized masker with -4 dB(A) difference between leaves noise and wind noise |

Figures 4.3, 4.4, and 4.5 show the mean obtained results for the speech intelligibility and subjective evaluation of the maskers at +10 dB(A), +13 dB(A), and +16 dB(A), respectively. Numerical values for these results can be seen in table 4.3.

As was expected, increasing the level difference between the masker and the speech signal resulted in a decreased intelligibility of the speech signal. However, it also resulted in a decrease in the perceived realness and pleasantness (i.e. comfort) of the masker, though it should be noted that the high standard deviation for these two parameters (almost always comparable in magnitude to the mean) makes it difficult to draw any clear conclusions about the obtained subjective measurements.

Analyzing the charts and table it also becomes apparent that the masking of female speech is more effective than that of male speech, particularly when using the synthesized maskers. In nearly all cases and levels, the synthesized maskers worked

twice as well for female speech as for male speech.

None of the synthesized maskers performed as well reducing the intelligibility of speech as the pink noise masker in any of the test situations. Table 4.4 shows the masking success (i.e. the number of mistakes provoked out of 25) as a percentage, highlighting the results for the pink noise masker and the best-performing masker in each other case. At the +16 dB(A) level difference the pink noise masker obtained near-perfect masking, with in an average of 23.08 mistakes (standard deviation = 1.55) for male speech (success rate of 92.33%) and 24.08 mistakes (standard deviation = 0.86) for female speech (success rate of 96.33%), whereas the best performing masker for male speech (*masker_m2*) resulted in an average of 10.83 mistakes (standard deviation = 3.00, success rate of 43.33%) and for female speech (*masker_0*) resulted in an average of 21.00 mistakes (standard deviation = 2.61, success rate of 84.00%). Also of note is the differences in standard deviation between the results from the pink noise masker and the synthesized ones at this level, which indicate a much more consistent masking success. Surprisingly, this is generally the opposite for other levels (see table 4.3).

Table 4.4 also helps to establish which of the synthesized maskers worked best. It evidences that for masking female speech, the masker in which there was no level difference between the leaves noise and the wind noise (*masker_0*) is the most appropriate, outperforming the others across all levels. However, the chosen masker should work well for masking both male and female speech. Because of this, *masker_m4* seems like the most appropriate, as it either has the highest masking success or is a close second in all but one of the situations. As discussed above, it should be noted that none of the maskers did a particularly good job at masking male speech and therefore the capacity to mask male speech should perhaps not be given as much weight. In any case, it seems clear that the maskers in which the leaves noise was higher than the wind noise (*masker_p2* and *masker_p4*) performed significantly worse than the others. This can probably be explained by the fact that most of the energy in speech is concentrated in the low and low-mid frequency range as seen in section 2.4, the region where the wind served as the primary masker.

Comparing the mistakes made at the three masker-speech level differences using the synthesized masker (figures 4.3a, 4.4a, and 4.5a) reveals that as the level difference between the masker and the speech increases, the leaves-sound to wind-sound masker level difference becomes less significant. In the +10 dB(A) difference case (4.3a), for female speech there is a difference of 4.25 mistakes between the worst-performing masker (*masker_p4*: 3.42 mistakes) and the best (*masker_0*, *masker_m2*, *masker_m4*: 7.67 mistakes), whereas for the +16 dB(A) difference case (4.5a) the difference is only 1.75 mistakes between the worst (*masker_m2*: 19.25 mistakes) and the best (*masker_0*: 21.00 mistakes).

Figure 4.6 shows a plot of of the total mistakes made for all maskers across the three different level differences with both the male (figure 4.6a) and female (figure 4.6b)

speech signals. These two plots seem to indicate is that there is close to a linear relationship between the level of the masker in comparison to the speech and the number of the mistakes made when using the synthesized masker.

Lastly, comparing the results of the subjective measurements of participant concentration, as seen in figure 4.7, suggests that the masker type or level was not a significant factor in affecting concentration, though again the high standard deviation of the results makes it difficult to draw any conclusions. In addition, several participants mentioned after taking the listening test that the intent of the question was not entirely clear and therefore results for this question should not be taken into account.

## 4.5 Test Limitations

It is important to acknowledge several known limitations of the modified Oldenburger Satztest that was carried out to evaluate the speech intelligibility of the synthesised maskers, that may or may not have affected the overall outcome of the results.

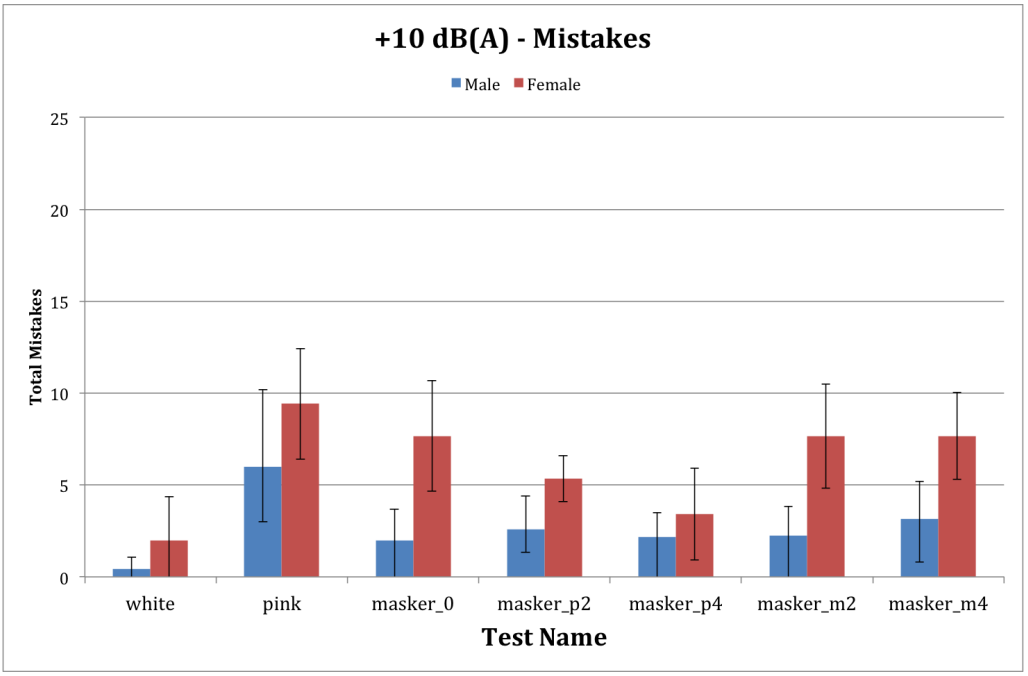
A key distinction between the test situation and the one that might be encountered in reality is that the masker and the speech were emitted from the same source (the loudspeaker system). In the complete, developed physical system, the masker would be emitted from a loudspeaker above the user, whereas the speech would be coming from some other point in the room, normally at a more horizontally comparable level. It is not known if this would have affected the outcome of the speech intelligibility, it should be considered. One way to test whether the masker and speech coming from the same source affects the intelligibility of speech would be to repeat the test using a configuration that more closely resembles the actual setup, with the masker coming from the above and the speech from somewhere in the horizontal plane and comparing the results.

In addition, the physical system being developed by the HS-OWL was not in place during the tests. This system, which includes the leaves and wind "source" that would act as the visual localizer of the masker noise, might have affected the perceived realness, if participants were able to visually identify the system as being the source of the noise.

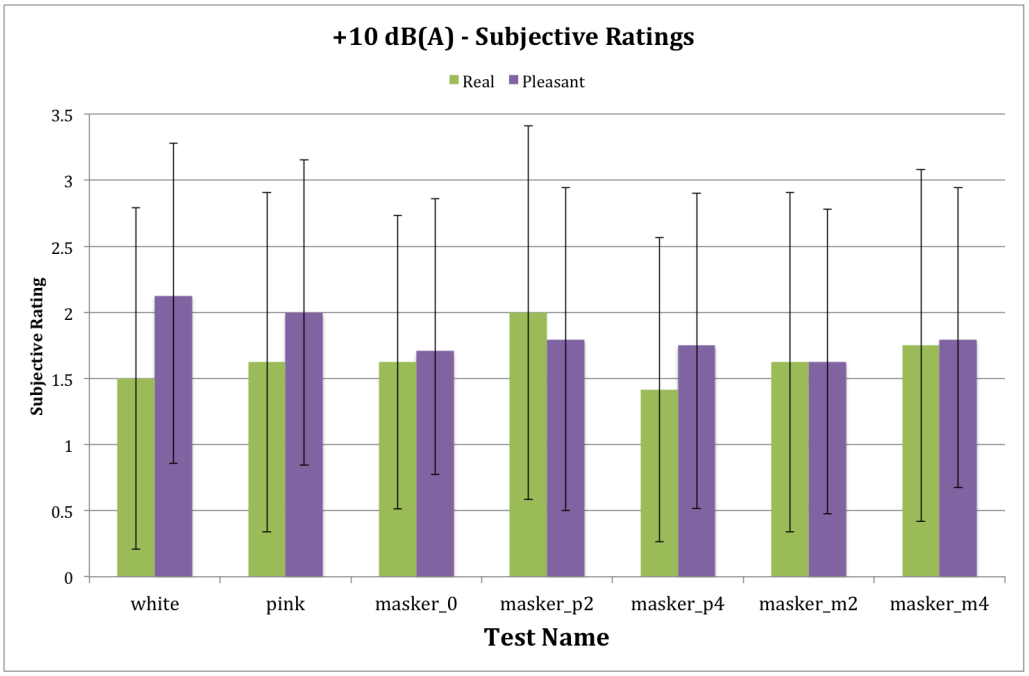
In the tests that were carried out, only one male voice and one female voice were used in the speech signals. As is known (and briefly discussed in section 2.4, each individual's vocal characteristics are different, and this test should not be used to make any conclusive statements about the overall performance of the masker with male and female speech. In order to have a better representation of speech characteristics, different male and female voices could have been used.

Another important difference between the designed test and the real office situation is that the test involved paying attention to the sound (speech + masker) in order to try and comprehend what the content of the speech signal was. The participants knew to expect speech coming through the noise, which lead to an increased effort to try and understand the words/sentences as opposed to accepting the noise as unintelligible and negligible. It is possible that as a result speech was more perceived to be more intelligible than if the user was carrying out some other task that did not require the comprehension of the emitted noise.

In order to analyze the effect of the partial masking as opposed to total masking, and to determine how the partially intelligible speech affects performance as described by the Irrelevant Speech Effect in section 1.1, it would be necessary to emit the maskers while participants undergo some kind of performance test. For example, a mental arithmetic test such as the one Schlittmeier et al. describe [36], in which test subjects have to carry out a series of mathematical operations under exposure to speech signals varying in intelligibility, could be used. Carrying out such a test would help to determine the "best" potential masker conditions, considering that Schlittmeier et al. conclude that "a combination of objective performance tests and subjective ratings is desirable for [a] comprehensive evaluation".

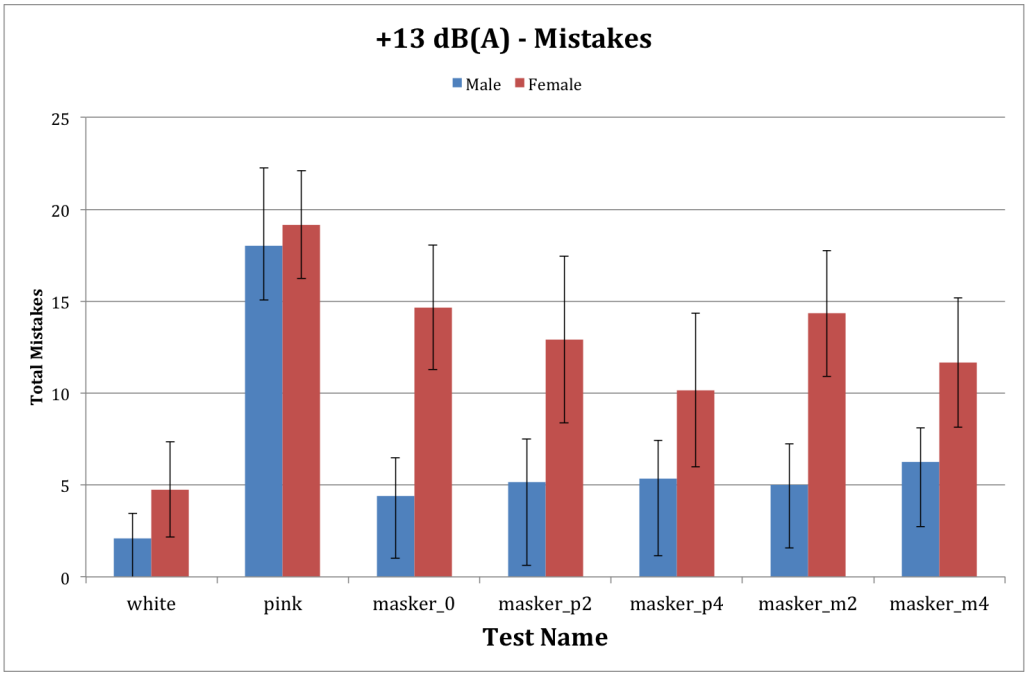



(a) Total number of mistakes

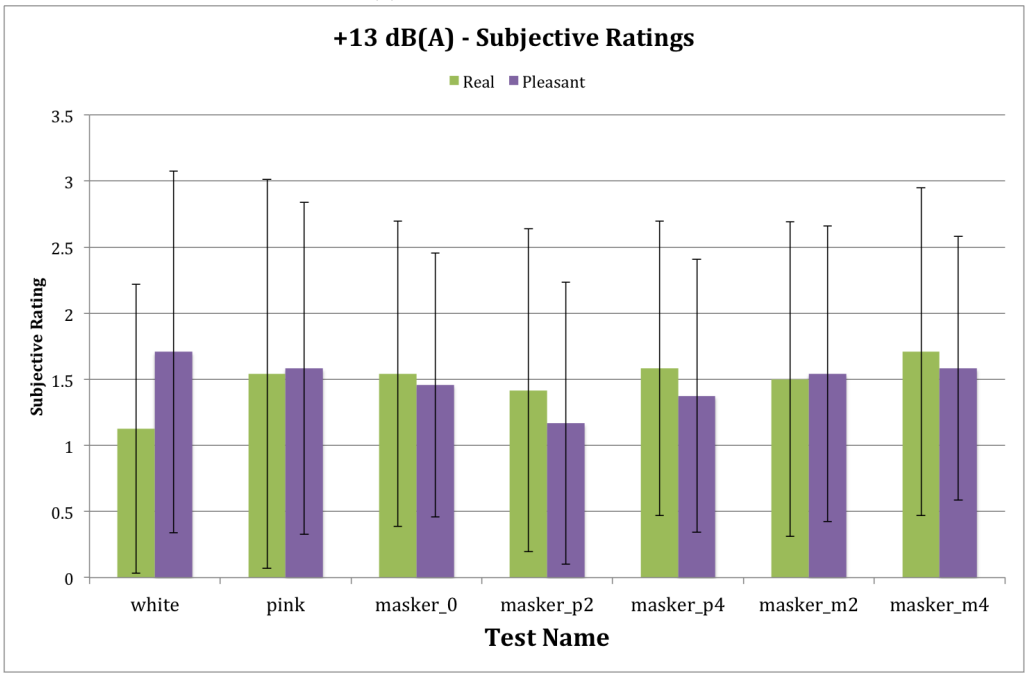



(b) Subjective realness and pleasantness ratings

Figure 4.3: Test results for the masker +10 dB(A) higher than the speech

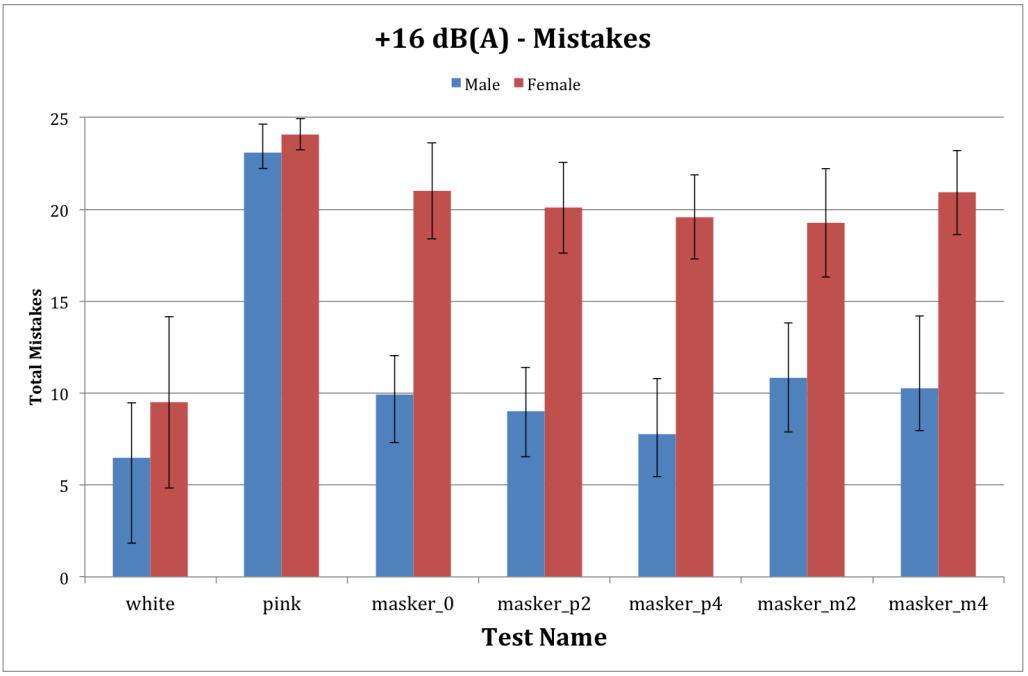



(a) Total number of mistakes

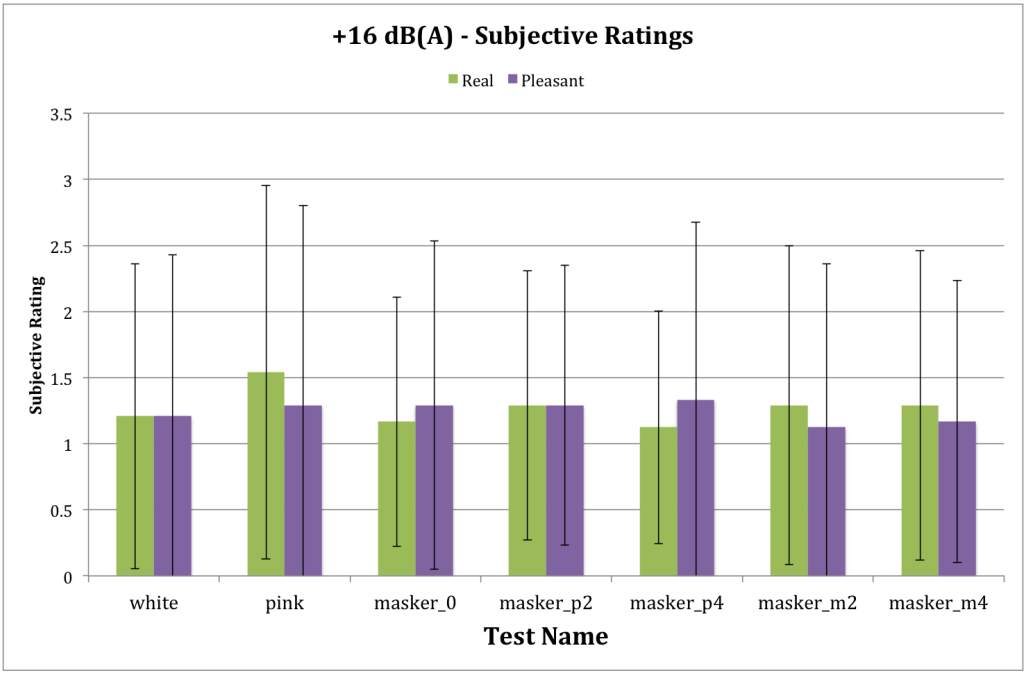



(b) Subjective realness and pleasantness ratings

Figure 4.4: Test results for the masker +13 dB(A) higher than the speech




(a) Total number of mistakes




(b) Subjective realness and pleasantness ratings

Figure 4.5: Test results for the masker +16 dB(A) higher than the speech

Table 4.3: Test results. Values given are means, with standard deviation in parentheses.

(a) Test results for tests for tests 'white', 'pink', and 'masker_0'

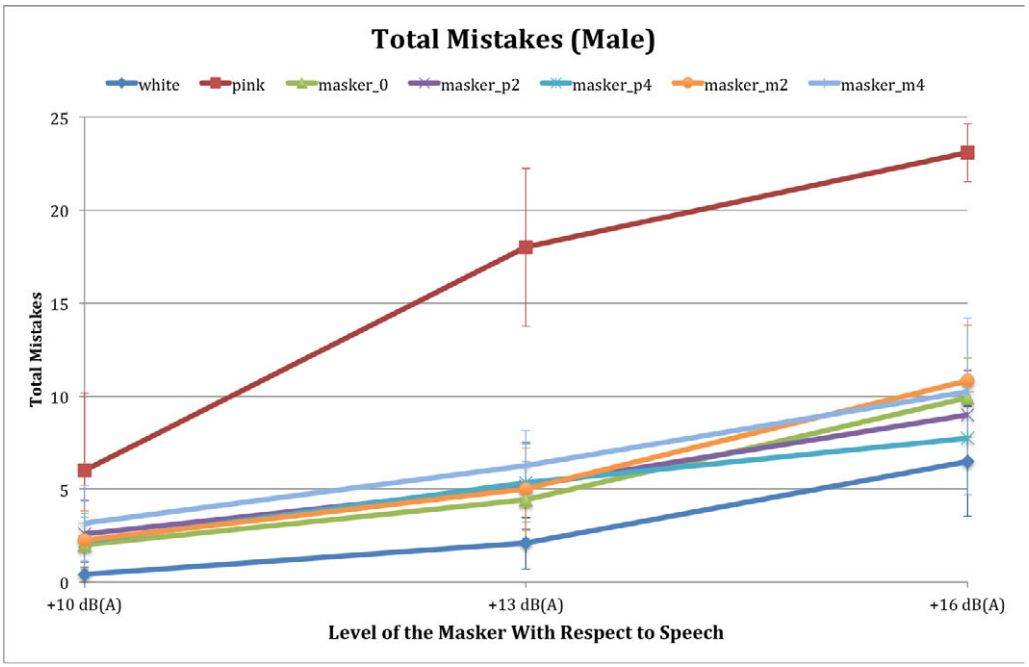
| Masker Level | Value | white | pink | masker_0 |
|---|---|---|---|---|
| +16 dB(A) | Mistakes Male | 6.50 (2.96) | 23.08 (1.55) | 9.92 (2.14) |
| | Mistakes Female | 9.50 (4.66) | 24.08 (0.86) | 21.00 (2.61) |
| | Real | 1.21 (1.15) | 1.54 (1.41) | 1.17 (0.94) |
| | Pleasant | 1.21 (1.22) | 1.29 (1.51) | 1.29 (1.24) |
| | Concentration | 3.83 (0.90) | 3.79 (1.00) | 3.83 (0.80) |
| +13 dB(A) | Mistakes Male | 2.08 (1.38) | 18.00 (4.24) | 4.42 (2.06) |
| | Mistakes Female | 4.75 (2.59) | 19.17 (2.94) | 14.67 (3.40) |
| | Real | 1.13 (1.09) | 1.54 (1.47) | 1.54 (1.15) |
| | Pleasant | 1.71 (1.37) | 1.58 (1.26) | 1.46 (1.00) |
| | Concentration | 3.96 (0.98) | 3.42 (0.91) | 3.71 (0.98) |
| +10 dB(A) | Mistakes Male | 0.42 (0.64) | 6.00 (4.18) | 2.00 (1.68) |
| | Mistakes Female | 2.00 (2.38) | 9.42 (3.01) | 7.67 (3.01) |
| | Real | 1.50 (1.29) | 1.63 (1.28) | 1.63 (1.11) |
| | Pleasant | 2.13 (1.27) | 2.00 (1.15) | 1.71 (0.93) |
| | Concentration | 3.54 (1.38) | 3.96 (0.93) | 3.79 (0.76) |

(b) Test results for tests for tests 'masker_p2', 'masker_p4', 'masker_m2', and 'masker_m4'

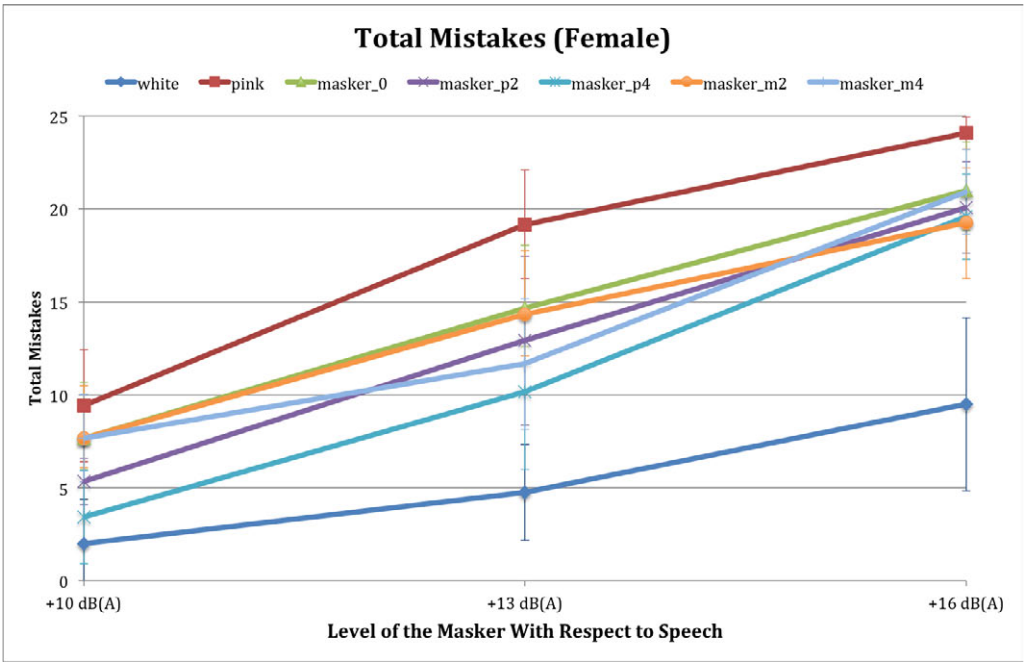
| Masker Level | Value | masker_p2 | masker_p4 | masker_m2 | masker_m4 |
|---|---|---|---|---|---|
| +16 dB(A) | Mistakes Male | 9.00 (2.38) | 7.75 (3.06) | 10.83 (3.00) | 10.25 (3.94) |
| | Mistakes Female | 20.08 (2.47) | 19.58 (2.29) | 19.25 (2.95) | 20.92 (2.29) |
| | Real | 1.29 (1.02) | 1.13 (0.88) | 1.29 (1.21) | 1.29 (1.17) |
| | Pleasant | 1.29 (1.06) | 1.33 (1.34) | 1.13 (1.24) | 1.17 (1.07) |
| | Concentration | 3.63 (0.90) | 3.54 (1.32) | 3.54 (0.87) | 3.46 (0.96) |
| +13 dB(A) | Mistakes Male | 5.17 (2.34) | 5.33 (2.09) | 5.00 (2.24) | 6.25 (1.88) |
| | Mistakes Female | 12.92 (4.54) | 10.17 (4.18) | 14.33 (3.42) | 11.67 (3.52) |
| | Real | 1.42 (1.22) | 1.58 (1.11) | 1.50 (1.19) | 1.71 (1.24) |
| | Pleasant | 1.17 (1.07) | 1.38 (1.03) | 1.54 (1.12) | 1.58 (1.00) |
| | Concentration | 3.46 (1.19) | 3.71 (1.06) | 3.63 (0.90) | 3.67 (0.85) |
| +10 dB(A) | Mistakes Male | 2.58 (1.80) | 2.17 (1.34) | 2.25 (1.59) | 3.17 (2.03) |
| | Mistakes Female | 5.33 (1.25) | 3.42 (2.50) | 7.67 (2.84) | 7.67 (2.36) |
| | Real | 2.00 (1.41) | 1.42 (1.15) | 1.63 (1.28) | 1.75 (1.33) |
| | Pleasant | 1.79 (1.29) | 1.75 (1.23) | 1.63 (1.15) | 1.79 (1.12) |
| | Concentration | 3.92 (1.08) | 3.63 (1.18) | 3.92 (1.19) | 3.75 (0.78) |

Table 4.4: Percentage of masking success for all maskers at +10 dB(A), +13 dB(A), and +16 dB(A) in comparison to speech. The percentage was calculated from the mean value shown in table 4.3. Highlighted in pink are the results for the pink noise masker and in green the best performing masker for each level.

| Masker Level | Parameter | white | pink | masker_0 | masker_p2 | masker_p4 | masker_m2 | masker_m4 |
|---|---|---|---|---|---|---|---|---|
| +16 dB(A) | Mistakes Male | 26.00 | 92.33 | 39.67 | 36.00 | 31.00 | 43.33 | 41.00 |
| | Mistakes Female | 38.00 | 96.33 | 84.00 | 80.33 | 78.33 | 77.00 | 83.67 |
| +13 dB(A) | Mistakes Male | 8.33 | 72.00 | 17.67 | 20.67 | 21.33 | 20.00 | 25.00 |
| | Mistakes Female | 19.00 | 76.67 | 58.67 | 51.67 | 40.67 | 57.33 | 46.67 |
| +10 dB(A) | Mistakes Male | 1.67 | 24.00 | 8.00 | 10.33 | 8.67 | 9.00 | 12.67 |
| | Mistakes Female | 8.00 | 37.67 | 30.67 | 21.33 | 13.67 | 30.67 | 30.67 |




(a) Male speech




(b) Female speech

Figure 4.6: Total number of mistakes for all maskers at +10 dB(A), +13 dB(A), and +16 dB(A) in comparison to speech. 4.6a shows results using male speech, and 4.6b shows results using female speech.

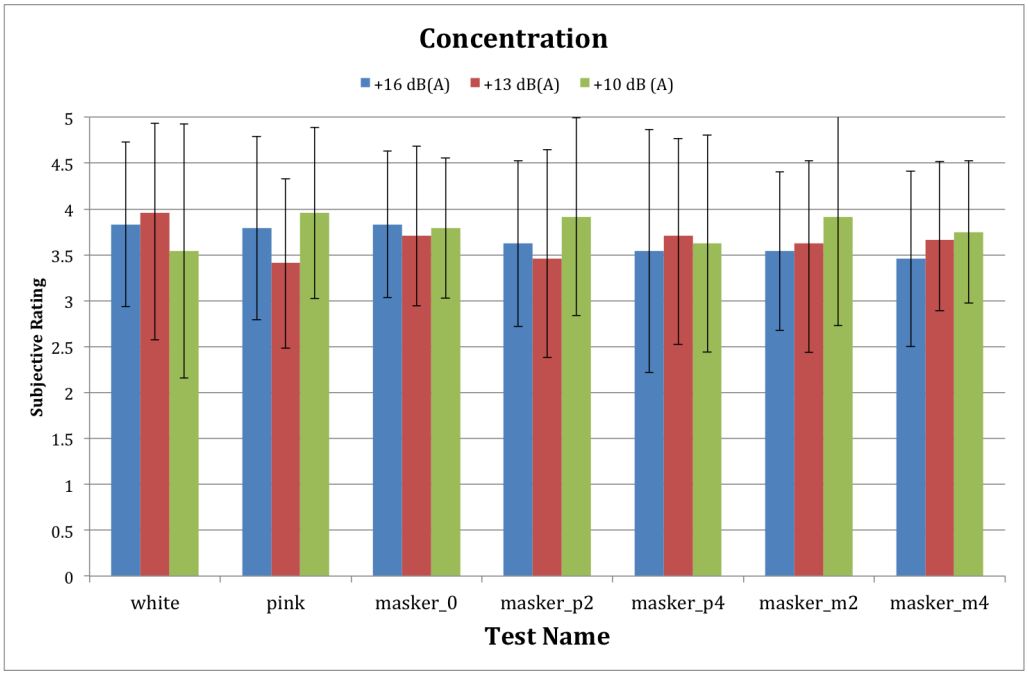



Figure 4.7: Test results for the subjective evaluation of concentration for the masker at +10 dB(A), +13 dB(A), and +16 dB(A) in comparison to speech.

# Chapter 5

# Conclusions and Future Work

## 5.1 Conclusions

The introduction of open-plan offices in the 1960s with the intent of encouraging flexibility, efficiency, and collaboration also caused a decrease in concentration and an increase in stress levels, among other negative consequences. Human speech in particular has proven to be an important source of these negative effects, and as such there has been an effort to find ways to reduce its intelligibility. One such way has been with the use of masking systems. However, to date the characteristics of the ideal masking noise have not been found. Though filtered noise has been used with moderate success, some studies suggest that natural noise maskers might be more successful in reducing the intelligibility of irrelevant background speech, as well as better accepted by the end users due to the possible visual localization of the source.

As part of the framework of a larger project with the goal of a coupled, adaptive noise masking system, in addition to a physical construct to be used as the apparent source, a natural noise was synthesized with the intent of being used as a speech masker. This speech masker would not only decrease the negative effect of irrelevant background speech on concentration and performance (Irrelevant Speech Effect), but also inherently increase speech privacy. In order to use the physical construct as a source of the sound, the synthesized noise was meant to simulate the sound of leaves rustling in the wind. Of course, the masker had to be generally accepted by the end users, who must not have found it to be too unpleasant or distracting.

Granular synthesis was used to synthesize the leaves noise, and subtractive synthesis was used to synthesize the wind noise. The noise was then tested for its effectiveness a modified version of an Oldenburger Satztest. In the test participants listened to a male or female speech signal consisting of five-word sentences and a masking noise and were expected to click on the correct words from a pre-given list to match the

sentence they believed to have heard. In addition, the test subjects were asked to subjectively evaluate the different masking noises according to the perceived "realness", pleasantness, and to rate their concentration. The maskers were emitted at three different levels with respect to the speech (+10 dB(A), +13 dB(A), +16 dB(A)), and seven different ones were used: white noise, pink noise, and five variations of the synthesized masker consisting of different proportions of leaves noise and wind noise. The goal of the test was not only to determine the masker's success at reducing speech intelligibility, but also to determine which of the variations of the synthesized maskers worked best, and at what level in comparison to the speech.

As was expected, the higher masker-to-speech level differences had a higher success at reducing the intelligibility of the speech signal for both male and female speech, reaching a maximum of 84.00% reduction in speech intelligibility, at a level 16 dB(A) higher than the speech signal, thus resulting in a significant increase in speech privacy. However, none of the synthesized maskers performed as well as the pink noise masker, which obtained a maximum success of 96.33%, at that same level (table 4.4, page 68). In all cases the maskers were more successful at masking the female speech than the male, and the difference was significant when using the synthesized maskers. Due to a high standard deviation in the results, the subjective evaluation of the perceived realness and pleasantness of the maskers proved to be inconclusive, though it's possible to infer that as the level of the masker with respect to speech increased, the subjective ratings for these two parameters decreased. The subjective evaluation of the concentration proved to be unsuccessful also due to a high standard deviation of the results, and confusion from the participants as to what the question was asking.

Though as the masker-to-speech level increased the overall differences between the results obtained for the different maskers decreased, of the synthesized maskers two seemed to be more successful, as described in section 4.4: *masker_0*, in which there was no level difference between the leaves noise and wind noise; and *masker_m4*, in which there was a -4 dB(A) difference between the leaves noise and the wind noise. Table 4.4 (page 68) showed the percentage of masking success for each of the maskers, revealing *masker_0* to be the most effective of the five synthesized ones masking female speech, and *masker_m4* to be masker that performed well in most cases.

The test results also seemed to indicate that the level difference between the leaves-noise and wind-noise became less important as the masker-to-speech level difference increased. Because of this, and given that in order to have a high (for example, above 60%) masking success rate the masker must be at least +13 dB(A) higher than the speech, the choice of masker between *masker_0* and *masker_m4* should not prove to be crucial. The subjective ratings for the perceived "realness" and "pleasantness" of each of the maskers are close enough that they also do not serve to distinguish between the two. However, due to *masker_m4*'s slightly better performance in

masking both male and female speech, it might be possible to conclude that it is the most successful of the five synthesized maskers.

Analyzing the spectral characteristics of the masker (figure 5.1), it is easy to observe a significant drop around the 1 kHz mark, where the wind-noise has been cut off. The spectral characteristics of this masker seem to be consistent with the findings of Veitch et al. [45] that effective maskers model the speech spectrum, which has less frequency content in the middle and high frequency range.

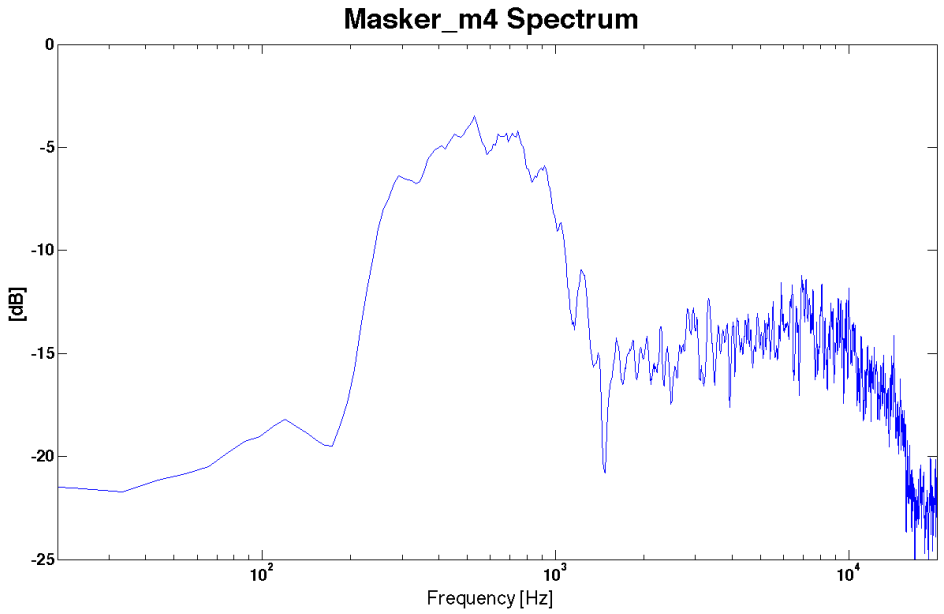



Figure 5.1: *Masker_m4* spectrum

The work done continues to collect information about the use of natural noises as speech maskers, a field that to date has had little published scientific research. The obtained results prove that a synthesized noise consisting of both leaves noise and speech noise could potentially serve as a masker. However, the levels at which it must be employed in comparison to the speech signal in order to have a high success rate might be too high. This high level difference might have been the cause of the apparent low user acceptance (via the perceived realness and pleasantness) of the masker, though as previously mentioned the results obtained in the subjective evaluation had a high standard deviation and are therefore not fully conclusive. Of the initial objectives established in section 1.3 (page 16, the user acceptance was the least well-fulfilled one, and will require further improvements.

## 5.2 Future Work

The proposed synthesized natural noise masker performed satisfactorily, but can still be improved in a few ways.

It was demonstrated that the synthesized noise masker performed significantly better when masking female speech than when masking male speech. This difference might be reduced by extending the frequency range of the masker in the lower frequencies. As discussed in section 2.4, the fundamental frequency of male speech can extend as low as 80 Hz, whereas the masker had a variable cutoff frequency centered at 250 Hz and with a variation of 125 Hz. It is possible that moving this cutoff frequency closer to 80 Hz would provide better masking, at the tradeoff of less "realistic" sounding wind.

The synthesized maskers proved to be generally not well accepted by the users, who rated it with a low score in both the perceived realness and pleasantness. While in part this may have been due to the high level at which they had to be emitted, it's also possible that the very constant leaf noise may have been the cause. Instead, it is proposed that the leaf-noise masker be synthesized with an even longer grain than the 150 to 300 millisecond ones used, and that the number of events (i.e. grains) be reduced from 500. While this may result in a less constant masker, the reduced presence and probable better "flow" should result in a more realistic and pleasant one. Of course, a middle point should be sought to balance between the synthesized noise's masking ability and the user acceptance.

It is also possible to use a different synthesis algorithm to synthesize either part of the speech masker. Schwarz and Schnell [40] propose two corpus-based concatenative synthesis methods of statistical modeling that allow for transitions between variations of the same texture. This method could be used to synthesize the wind-noise masker, for instance, allowing for instances of light to heavy wind according to the level of the masker. The wavelet-based approach used by O'Regan and Kokaram [29] could be used as s an alternative to subtractive synthesis for synthesizing the leaf-noise. Their method, based on an algorithm for image texture synthesis, achieves a large segment size that is well adapted to the source, which could provide the necessary realness for the semi-random rustling of the leaves in the wind.

As already discussed in section 1.1, irrelevant speech and/or certain non-speech sounds can impair short-term memory performance. One of the objectives of the synthesized masker, as described in section 1.3, was that it did not disrupt concentration. However, the tests that were carried out only objectively evaluated the masker's ability to impair speech intelligibility, and subjectively evaluated its perceived realness and pleasantness. In order to determine how the maskers affect performance, it would be necessary to emit them while participants undergo some kind of performance test. For example, a mental arithmetic test such as the one Schlittmeier et al. [36] describe, in which test subjects have to carry out a series of

mathematical operations under exposure to speech signals varying in intelligibility, could be used.

In addition, it might be of use to perform a more refined version of the speech intelligibility test, using smaller intervals between masker-to-speech level differences. For example 1 dB(A) intervals from a +13 dB(A) level difference to a +16 dB(A) level difference could be used, since these levels had a masking success rate of over 50%, as seen in in table 4.4 on page 68. If this refined version of the speech intelligibility test is carried out alongside a mental arithmetic test such as the one described above, an ideal middle point between performance and masking could be found.

Lastly, it would be good to see if the developed physical structure that simulates the masker source affects the results obtained in the subjective evaluation, perhaps by increasing the perceived realness.

# Appendices

# Appendix A

# Pitch-Shifting Algorithm

As explained in section 3.3.1 (page 46), the pitch-shifting algorithm was taken from Grondin [14].

Pitch-shifting consists of scaling a frequency, or group of frequencies, down or up by a certain factor. This can be seen in (A.1), where $s$ is the factor to be pitch-shifted. One possibility to do this is to resample the original signal at some new sampling frequency (namely, $f_{original} \cdot s$). However, this results in a duration of the signal that is $L/s$ as long as the original (see figure A.1), which is not desired.

$$f_{\text{shifted}} = f_{\text{original}} \cdot s \tag{A.1}$$

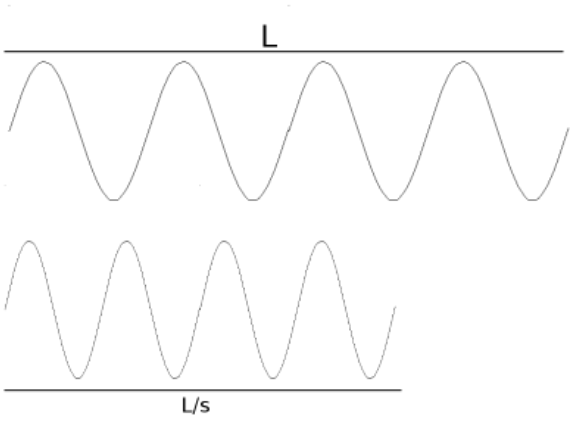



Figure A.1: Result of pitch-shifting by resampling, where $s$ is the pitch-shift factor

Instead, a better method is to effectively double the length of the original signal (to $L \cdot s$) without affecting the pitch and then resampling so as to obtain a pitch-shifted signal of the original duration $L$.

In order to do so, the signal will first be split into several overlapping frames. These frames will then be either spaced further apart (to stretch) or closer together (to compress) in order to create the time-stretched signal that will then be resampled, as shown in figure A.2. This new spacing, ($s$ times the original), however, leads to discontinuities in the signal, which may be heard as glitches. To resolve this, it is necessary to compensate for the phase shifts.

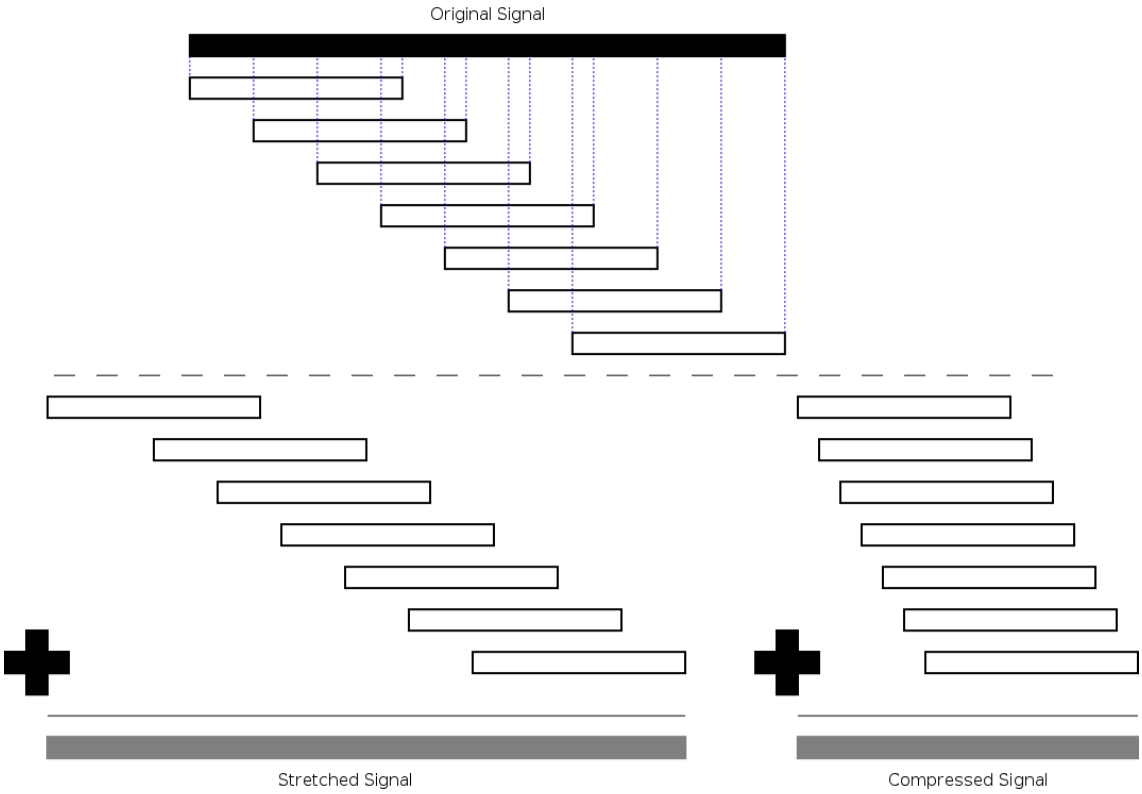



Figure A.2: Separation into frames, and resulting stretched or compressed signals

Before calculating phase differences, it is first necessary to transform to the frequency domain. To do so, the frame is first windowed using a Hanning window, and then transformed using a Fast Fourier Transform (FFT). This equation is shown in (A.2), where $x[n]$ is the original signal, $w[n]$ is the Hanning window, and $(X_a[k])_i$ [1] is the discrete spectrum of frame $i$. $Step_a$ is the number of samples between two successive windows and in this case is equal to $\frac{N}{4}$ since a 75% overlap is used.

$$(X_a[k])_i = \sum_{n=0}^{N-1} x[n + i * (\text{step}_\text{a})] \cdot w[n] e^{-j(\frac{2\pi kn}{N})} \qquad k = 0, 1, 2, ..., N-1 \qquad \text{(A.2)}$$

The application of an N-length FFT with a sampling frequency of $f_s$ results in having a symmetrical spectrum and N frequency bins from 0 to $\frac{(N-1)}{N} f_s$ with a resolution of

[1] From here onwards $k$ will be refer to a bin index and $i$ will refer to a frame index

$\frac{f_s}{N}$. If a signal has a frequency that falls between two bins, its energy will be spread out to the nearby bins.

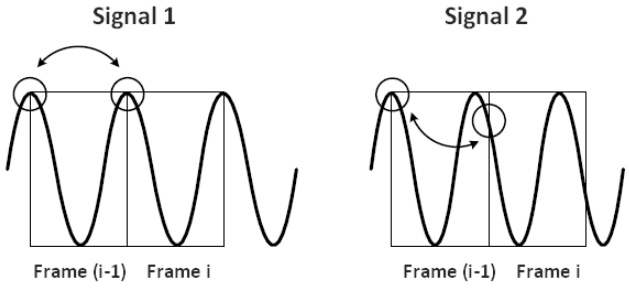



Figure A.3: Sine waves with different frequencies and a phase difference, from [14]

Figure A.3 shows two sine waves of different frequencies, split into frames of $N$ samples without overlap (for simplicity). The first sine wave has a frequency of $\frac{f_s}{N}$ and therefore falls exactly in the first bin. The second has a frequency that is slightly higher, and therefore doesn't fall exactly in the first bin, though it does the bin does contain most of its energy. In the case of the first signal, there is no phase difference between the two consecutive frames; in the case of the second, the phase difference is greater than zero, which corresponds to a signal of a frequency higher than the bin frequency. This phase difference is known as a phase shift, $(\Delta\phi_a[k])_i$, and is used to determine the true frequency associated with the bin. This phase information is warped between $-\pi$ and $\pi$, however, and must first be unwrapped.

Equations (A.3), (A.4), and (A.5) illustrate the method of unwrapping the phase, consisting of first calculating and wrapping the frequency deviation from the bin, and then adding this to the bin frequency. In the equations $\Delta t_a$ is the time interval between two frames ($\text{step}_\text{a}$ divided by $f_s$), $w_{\text{bin}}[k]$ is the bin frequency, $(\Delta w[k])_i$ is the frequency deviation of the current frame, and $(\Delta w_{\text{wrapped}}[k])_i$ is the wrapped frequency deviation.

$$(\Delta w[k])_i = \frac{(\phi_a[k])_i - (\phi_a[k])_{i-1}}{\Delta t_a} - w_{\text{bin}}[k] \tag{A.3}$$

$$(\Delta w_{\text{wrapped}}[k])_i = \text{mod}[((\Delta w[k])_i + \pi), 2\pi] - \pi \tag{A.4}$$

$$(w_{\text{true}}[k])_i = w_{\text{bin}}[k] + (\Delta w_{\text{wrapped}}[k]_i) \tag{A.5}$$

By multiplying the true frequency obtained in (A.5) with the time interval of the desired signal (i.e. the pitch-shifted one), the new phase can be calculated, as shown in (A.6). In this equation $\Delta ts$ is the step size of the pitch-shifted signal $\text{step}_\text{s}$, equivalent to the multiplication of the original step size $\text{step}_\text{a}$ with the pitch-shift

factor $s$. The phase from the previous frame of the synthesis is known, as it was already calculated in the previous iteration of the algorithm.

$$(\phi_s[k])_i = (\phi_s[k])_{i-1} + \Delta t_s \cdot (w_{\text{true}}[k])_i \tag{A.6}$$

Finally the new spectrum is obtained as illustrated in equation (A.7).

$$|(X_s[k])_i| = |(X_a[k])_i| \quad \angle(X_s[k])_i = (\phi_i)_s \tag{A.7}$$

After adjusting the phase the next step is to revert back to the time domain and overlap-add the frames together. For the former an Inverse Discrete Fourier Transform (IDFT) is used, which is then windowed using a Hanning window in order to smooth the signal, as shown in equation (A.8). Equation (A.9) shows the process for overlap-adding the frames back together, where $L$ is the frame number and $u[n]$ is the unit step function.

$$q_i[n] = \left\{ \frac{1}{N} \sum_{k=0}^{N-1} (X_s[k])_i e^{-j(\frac{2\pi k n}{N})} \right\} \cdot w[n] \quad n = 0, 1, 2, ..., N-1 \tag{A.8}$$

$$y[n] = \sum_{i=0}^{L-1} q_i[n - i \cdot \text{step}_\text{s}] \cdot \{u[n - i \cdot \text{step}_\text{s}] - u[n - i \cdot \text{step}_\text{s} - N]\} \tag{A.9}$$

This process results in a signal that is either stretched or compressed in time a factor of $s$, without any variation in pitch. In order to obtain the final, pitch-shifted signal, a resampling at $f_s \cdot s$ is performed, using linear interpolation if needed to approximate the desired samples.

# Appendix B

# Graphic User Interface (GUI) for the Speech Intelligibility Test

The following figures show the developed graphic interface for the speech intelligibility test described in chapter 4. All of the GUIs were made using MATLAB.


Figure B.1: Test subject data input

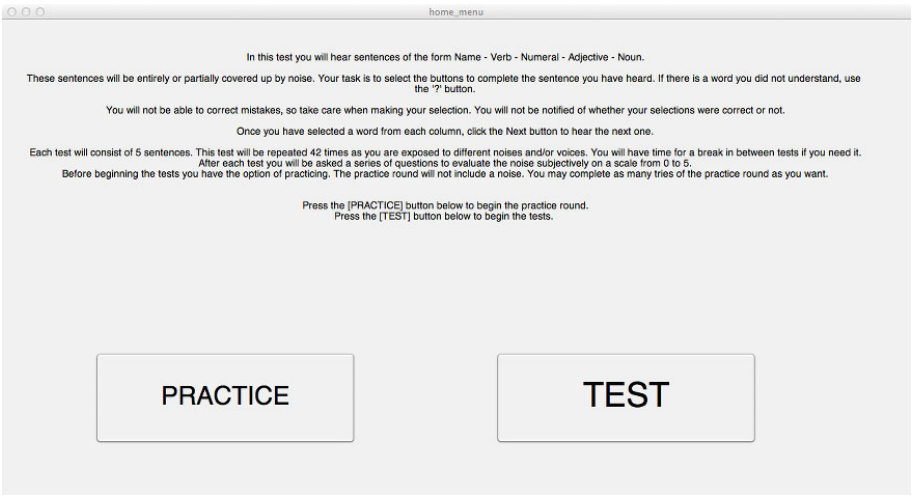



Figure B.2: Home Menu

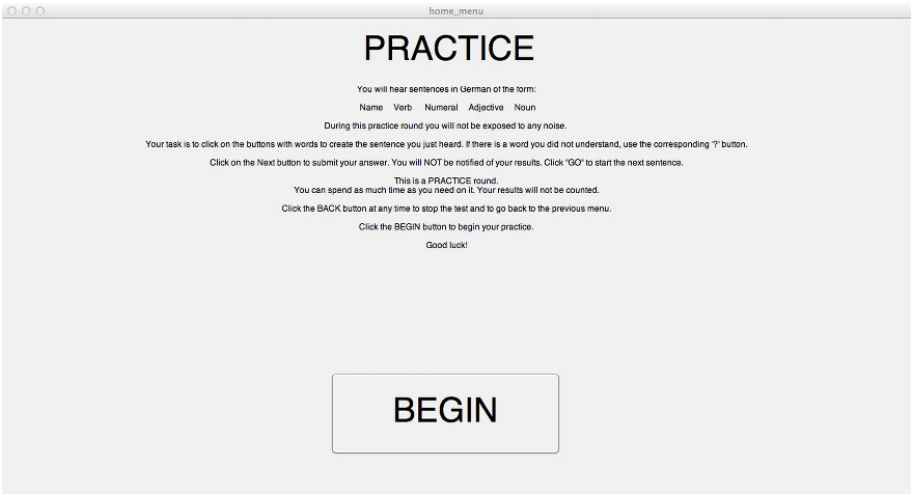



Figure B.3: Instructions given for the practice round

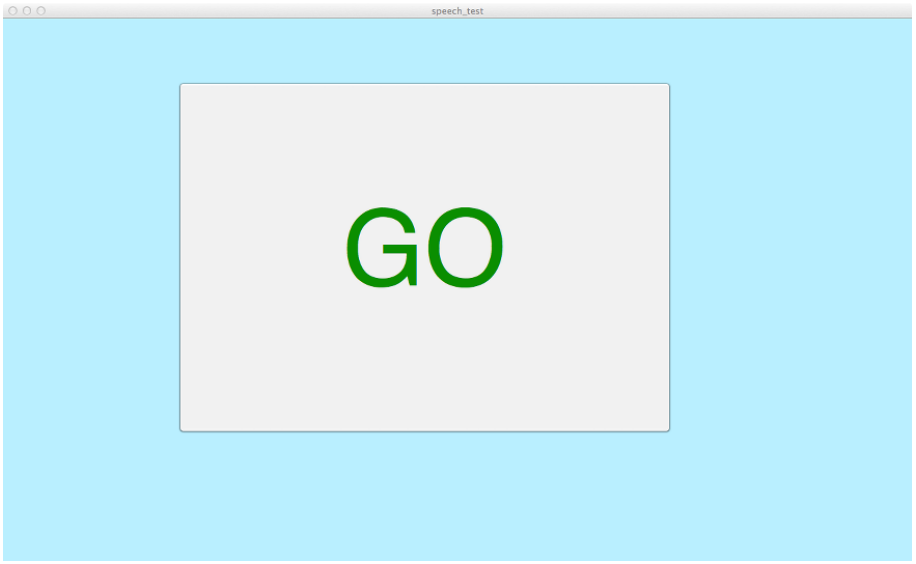



Figure B.4: Go button that started the tests


Figure B.5: Graphic interface for the practice round

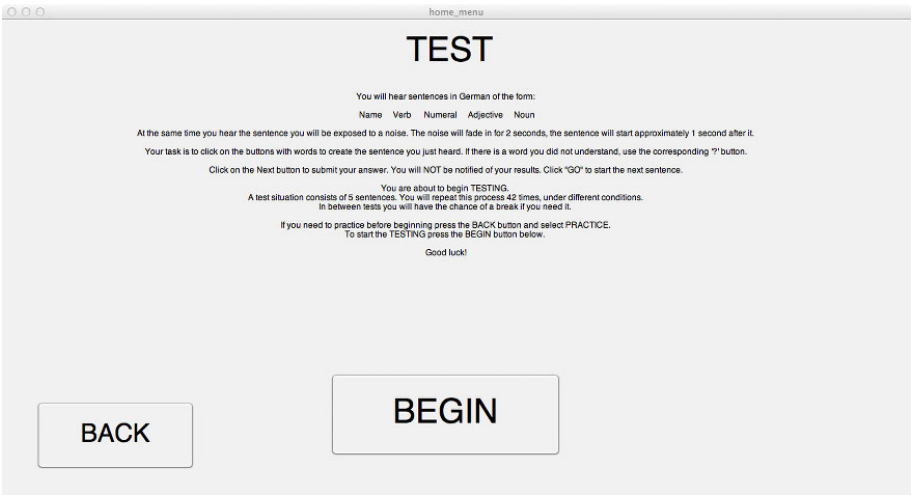



Figure B.6: Instructions given for the test

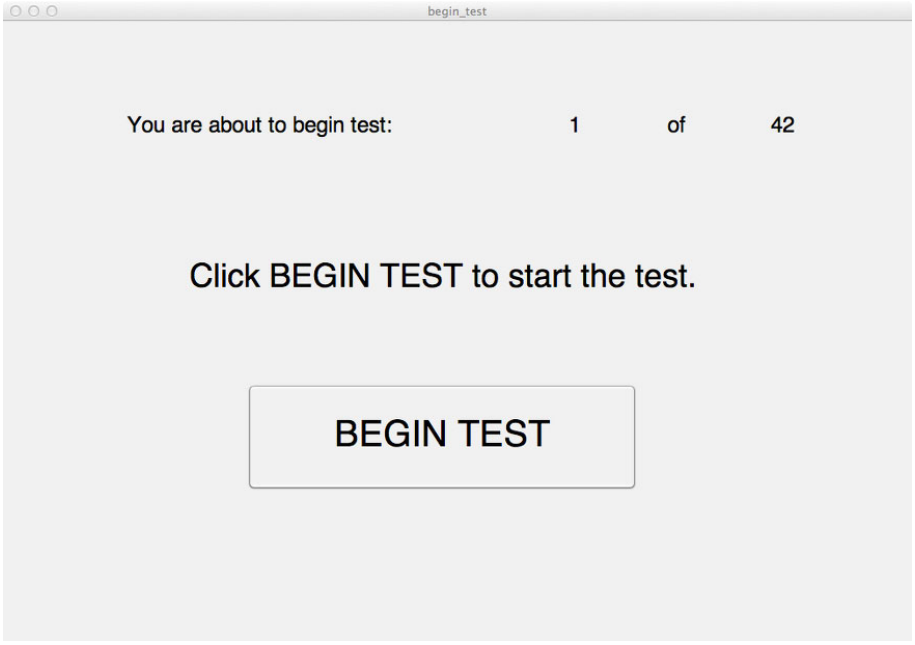



Figure B.7: Screen notifying the test number that was about to be started


Figure B.8: Graphic interface for the test

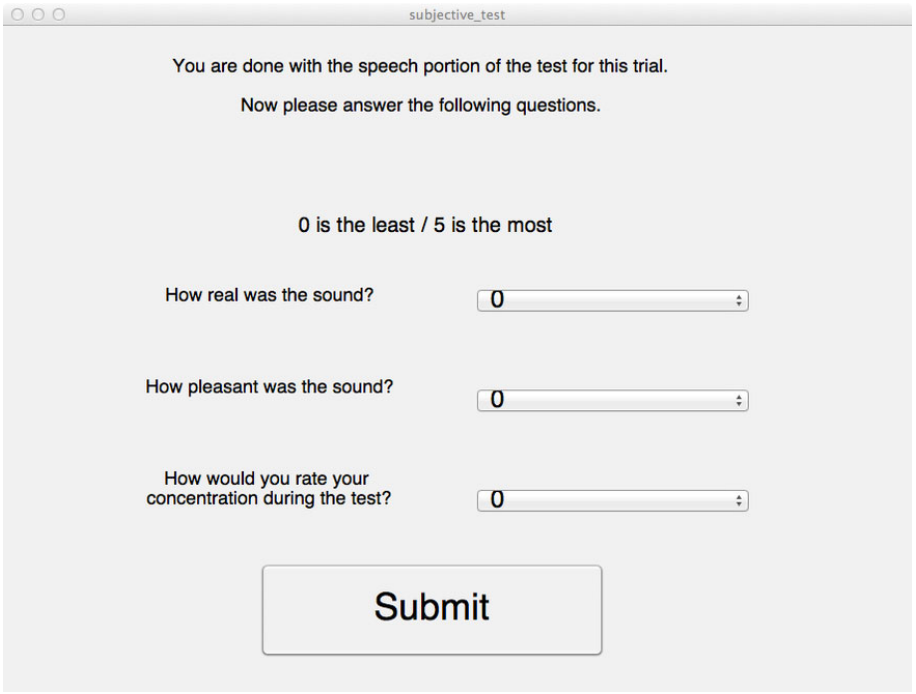



Figure B.9: Graphic interface for the subjective evaluation

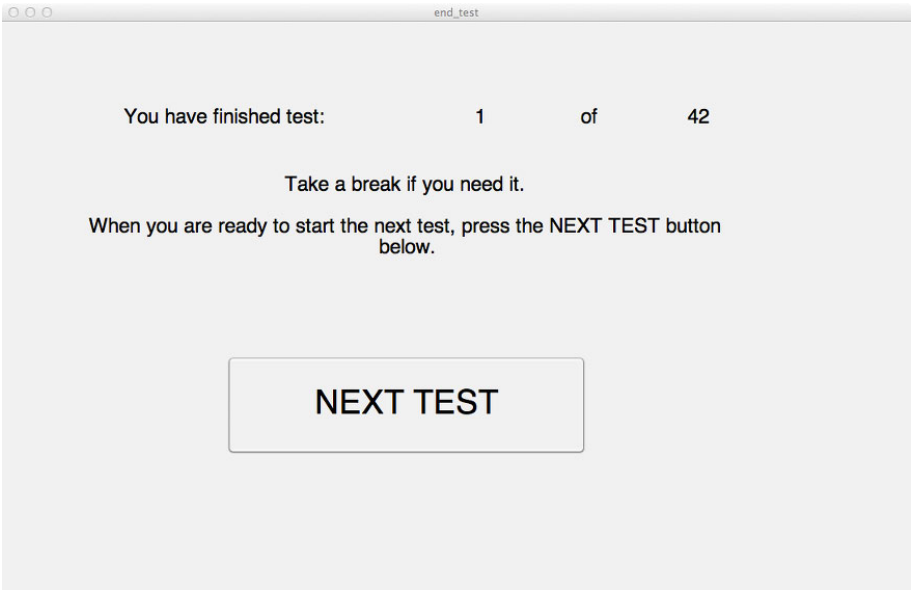



Figure B.10: Screen notifying the test number that was just completed

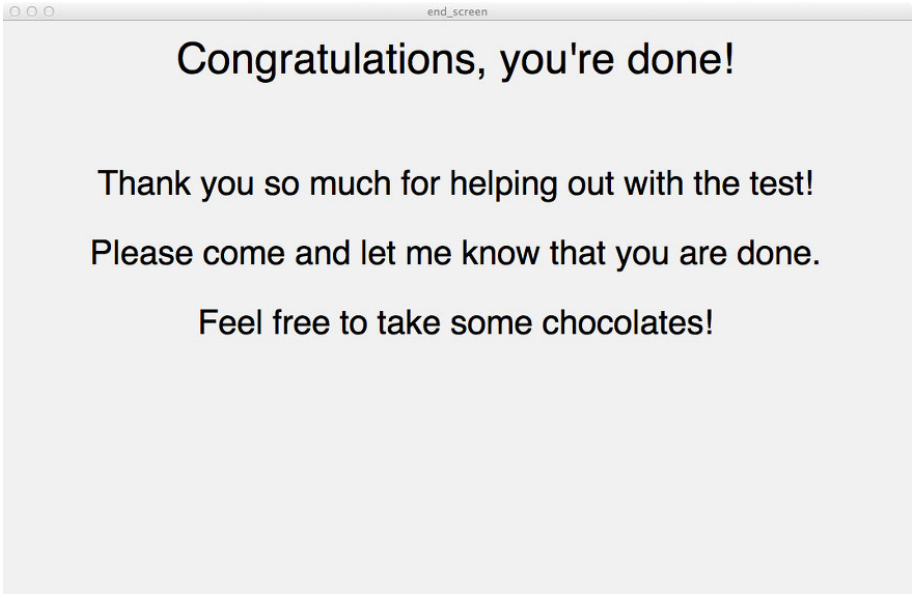



Figure B.11: Screen that the test had been completed

# Bibliography

[1] K. Abe, K. Ozawa, Y. Suzuki, and T. Sone. Comparison of the effects of verbal versus visual information about sound sources on the perception of environmental sounds. *Acta Acustica united with Acustica*, 92:51–60, 2006.

[2] J. B. Allen. Nonlinear cochlear signal processing and masking in speech perception. *Springer Handbook of Speech Processing*, pages 27–60, 2008.

[3] S. Banbury and D. C. Berry. Disruption of office related tasks by speech and office noise'. *British Journal of Psychology*, 89(3):499–517, August 1998.

[4] M. Bastiaans. Gabor's expansion of a signal into gaussian elementary signals. *Proceedings of the IEEE*, 68:538–539, 1980.

[5] J. Blauert and N. Xiang. *Acoustics for Engineers*. Springer, 2008.

[6] J. Bradley and B. Gover. Criteria for acoustic comfort in open-plan offices. *Proceedings of the Inter-noise 2004, the 33rd International Congress and Exposition on Noise Control Engineering*, pages 1–6, August 2004.

[7] P. Cook. Physically informed sonic modeling (PhISM): Percussive synthesis. *Proceedings of the International Computer Music Conference (ICMC)*, 21(3):38–49, 1996.

[8] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman. Synthesizing sound textures through wavelet tree learning. *Virtual Words, Real Sounds*, July/August 2002.

[9] G. W. Evans and D. Johnson. Stress and open-office noise. *Journal of Applied Psychology*, 85(5):779–783, 2000.

[10] G. Fant. *Acoustic Theory of Speech Production*. Mouton & Co., The Hague, 1960.

[11] H. Fastl and E. Zwicker. *Psychoacoustics: Facts and Models*. Springer, third edition, 2007.

[12] D. Gabor. Acoustical quanta and the theory of hearing. *Nature*, 159:591–594, 1947.

[13] C. Gil González. *Introducción a las Salas Para la Palabra*. Escuela Universitaria de Ingeniería Técnica de Telecomunicación, Universidad Politécnica de Madrid, Madrid, 2003.

[14] F. Grondin. Guitar pitch shifter, 2009.

[15] A. Haapakangas, E. Kankkunen, V. Hongisto, P. Virjonen, D. Oliva, and E. Keskinen. Effects of five speech masking sounds on performance and acoustic satisfaction. implications for open-plan offices. *Acta Acustica united with Acustica*, 97:641–655, 2011.

[16] J. R. Hanley and E. Bakopoulou. Irrelevant speech, articulatory suppression, and phonological similarity: A test of the phonological loop model and the feature model. *Psychonomic Bulletin & Review*, 10(2):435–444, 2003.

[17] H. C. Hardy. A guide to office acoustics. *Architectural Record*, 121(2):235–240, 1957.

[18] V. Hongisto. Effect of sound masking on workers in an open office. *Proceedings of Acoustics 08 Paris*, pages 537–542, 2008.

[19] HörTech gGmbh, Oldenburg, Germany. *Oldenburger Satztest: Adaptive Sprachaudiometrie mit Sätzen in Ruhe und im Störgeräusch*, 2011.

[20] R. Hoskinson and D. K. Pai. Manipulation and resynthesis with natural grains. *Proceedings of the International Computer Music Conference (ICMC)*, 2001.

[21] A. Huovilainen and V. Välimäki. New approaches to digitial subtractive synthesis. *Proceedings of the International Computer Music Conference (ICMC)*, pages 399–402, September 2005.

[22] D. M. Jones, C. Miles, and J. Page. Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory. *Applied Cognitive Psychology*, 4:89–108, 1990.

[23] H. Lane and B. Tranel. The Lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, 14:677–709, December 1971.

[24] D. D. Marshall. Multimedia module no: Cm0340. Online Resource, 2013.

[25] A. Misra and P. Cook. Toward synthesized environments: A survey of analysis and synthesis methods for sound designers and composers. *Proceedings of the International Computer Music Conference (ICMC)*, pages 155–162, 2009.

[26] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Elsevier Academic Press, London, fifth edition, 2004.

[27] J. Mullen. *Physical Modelling of the Vocal Tract with the 2D Digital Waveguide Mesh*. PhD thesis, The University of York, England, 2006.

[28] J. Nemecek. Music during office work. In E. Grandjean, editor, *Ergnomics and Health in Modern Offices*, pages 64–69. Taylor and Francis, London, 1984.

[29] D. O'Regan and A. Kokaram. Multi-resolution sound texture synthesis using the dual-tree complex wavelet transform. *Proceedings of the 15th European Signal Processing Conference*, pages 350–354, September 2007.

[30] V. H. P. Virjonen, J. Keränen. Determination of acoustical conditions in open-plan offices. Proposal for new measurement method and target values. *Acta Acustica united with Acustica*, 95:279–290, 2009.

[31] M. Park, A. Kohlrausch, and A. van Leest. Irrelevant speech effect under stationary and adaptive masking conditions. Unpublished, July 2013.

[32] K. Pohlmann. *Principles of Digital Audio*. McGraw-Hill, sixth edition, 2010.

[33] C. Roads. Introduction to granular synthesis. *Computer Music Journal*, 12(2):11–13, Summer 1988.

[34] M. Russ. Physical modelling synthesis explained. *Sound on Sound*, June 1997.

[35] N. Saint-Arnaud and K. Popat. Analysis and synthesis of sound textures. *Readings in Computational Auditory Scene Analysis*, pages 125–131, 1995.

[36] S. Schlittmeier, J. Hellbrück, R. Thaden, and M. Vorländer. The impact of background speech varying in intelligibility: Effects on cognitive performance and perceived disturbance. *Ergonomics*, 51(5):719–736, May 2008.

[37] M. Schmitt. Projekttreffen zum ZIM-Projekt "Private Workspace": Bericht ETI. (Personal Communication), December 2013.

[38] D. Schwarz. Concatenative synthesis: The early years. *Journal of New Music Research*, 35(1):3–22, 2006.

[39] D. Schwarz. State of the art in sound texture synthesis. *Proceedings of the 14th Int. Conference on Digital Audio Effects (DAFx-11)*, pages 19–23, 2011.

[40] D. Schwarz and N. Schnell. Descriptor-based sound texture sampling. *Proceedings of the International Conference on Sound and Music Computing (SMC) Conference*, pages 510–515, July 2010.

[41] D. Simón Zorita. *Análisis y síntesis de señales musicales*. Escuela Universitaria de Ingeniería Técnica de Telecomunicación, Universidad Politécnica de Madrid, 2000.

[42] G. Strobl, G. Eckel, and D. Rocchesso. Sound texture modeling: A survey. *Proceedings of the International Conference on Sound and Music Computing (SMC)*, 2006.

[43] I. R. Titze. *Principles of Voice Production*. Prentice Hall, 1994.

[44] K. van den Doel, P. G. Kry, and D. K. Pai. Foleyautomatic: Physically-based sound effects for interactive simulation and animation. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 537–544, 2001.

[45] J. Veitch, J. Bradley, L. Legault, S. Norcross, and J. Svec. Masking speech in open-plan offices with simulated ventilation noise: Noise level and spectral composition effects on acoustic satisfaction. Technical report, Institute for Research in Construction, Ontario, Canada, 2002.

[46] N. Venetjoki, A. Kaarlela-Tuomaala, E. Keskinen, and V. Hongisto. The effect of speech and speech intelligibility on task performance. *Ergonomics*, 49:1068–1091, 2006.

[47] C. Verron. *Synthèse immersive de sons d'environnement*. PhD thesis, Universitè de Provence, 2010.

[48] C. Verron, M. Aramaki, R. Kronland-Martinet, and G. Pallone. Analysis/synthesis and spatialization of noisy environmental sounds. *Proceedings of the 15th International Conference on Auditory Display*, pages ICAD09–1–ICAD09–5, May 2009.

[49] R. A. Waller. Office acoustics. effect of background noise. *Applied Acoustics*, 2:121–130, 1969.

[50] J. Watkinson. *The Art of Digital Audio*. Focal Press, third edition, 2001.

[51] I. Xenakis. *Formalized Music*. Indiana University Press, Bloomington, 1971.

[52] F. Zickmantel, G. Papsdorf, M. Kob, and U. Pottgiesser. Private Workspace: Entwicklung eines adaptiven Schallmaskierungssystems für offene Arbeitsbereiche. (Personal Communication), 2012.

# *Eigenständigkeitserklärung*
# Statement

*Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig verfasst habe. Ich versichere, dass ich keine anderen als die angegebenen Hilfsmittel und Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe, und dass die eingereichte Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahren gewesen ist.*

I certify that this Bachelor's thesis is the product of my own work. I confirm that I did not use any sources or resources other than those specified, marked all statements verbally or analogously taken from other works as such, and that this submitted thesis has neither been all nor in significant parts subject to another examination procedure.

Detmold, 8 Juli 2014

Jorge Mir Álvarez